

Maximum Likelihood Estimation and Forecasting for GARCH, Markov Switching, and Locally Stationary Wavelet Processes

Yingfu Xie

*Centre of Biostochastics
Department of Forest Economics
Umeå*

**Doctoral Thesis
Swedish University of Agricultural Sciences
Umeå 2007**

Acta Universitatis Agriculturae Sueciae
2007:107

ISSN 1652-6880
ISBN 978-91-85913-06-0
© 2007 Yingfu Xie, Umeå
Printed by: Arkitektkopia, Umeå 2007

Abstract

Yingfu Xie. *Maximum Likelihood Estimation and Forecasting for GARCH, Markov Switching, and Locally Stationary Wavelet Processes*. Doctoral Thesis.

ISSN 1652-6880, ISBN 978-91-85913-06-0.

Financial time series are frequently met both in daily life and the scientific world. It is clearly of importance to study the financial time series, to understand the mechanism giving rise to the data, and/or predict the future values of a series. This thesis is dedicated to statistical inferences of a number of models for financial time series.

Financial time series often exhibit time-varying and clustering volatility (conditional variance), which were not handled well by traditional models, until the development of the autoregressive conditionally heteroscedastic (ARCH) and the generalized ARCH (GARCH) models. We prove the consistency and asymptotic normality of the quasi-maximum likelihood estimators for a GARCH(1,2) model with dependent innovations, which extends the results for the GARCH(1,1) model in the literature under weaker conditions.

The regime-switching GARCH (RS-GARCH) model extends the GARCH models by incorporating a Markov switching into the variance structure. The statistical inferences for the RS-GARCH model are difficult due to the complex dependence structure. One alternative is to take average over all regimes at every step, and adapt the integrated conditional variances. Another one is to transform the GARCH into an ARCH model. The maximum likelihood (ML) estimation of these two cases is considered. Consistency of the ML estimators is proved, and the asymptotic normality is suggested by simulation studies. The results are further generalized to a general autoregressive model with Markov switching, in which the autoregression can be of infinite order. Consistency of the ML estimators is obtained and the asymptotic normality is conjectured.

Time series analysis can also be conducted in frequency domain, i.e. to analyze their spectral values obtained by e.g. Fourier or wavelet transforms. Locally stationary wavelet (LSW) processes are a class of processes defined on a set of non-decimated wavelets. We first address the problem on how to select a wavelet in practice, and some guidelines are suggested by simulation studies. The existing forecasting algorithm for LSW processes is found vulnerable to outliers, and a new forecasting algorithm is proposed to overcome this weakness. The new algorithm is shown stable and outperforms the existing algorithm when applied to real financial data. The volatility forecasting ability of LSW model based on our new algorithm is then discussed and is shown to be competitive with GARCH models.

Algorithms and functions for data generation, calculation and maximization of the likelihoods for RS-GARCH models and the new forecasting algorithm of LSW processes are appended.

Key words: Consistency; Financial time series; Forecasting; GARCH; LSW process; Maximum likelihood estimation; Markov switching; Non-decimated wavelet; Volatility forecasting

Author's address: Centre of Biostochastics, SLU, 901 83 Umeå, Sweden.

Contents

Maximum Likelihood Estimation and Forecasting for GARCH, Markov Switching, and Locally Stationary Wavelet Processes

1	Introduction	1
2	Heteroscedastic time series and GARCH models	6
2.1	The GARCH model	7
2.2	Basic properties of the GARCH model	8
2.3	Quasi-MLE	9
2.4	Asymptotic properties	10
2.5	Other models	12
3	Markov switching models	13
3.1	Regime-switching GARCH and path dependence problem	14
3.2	The reduced RS-GARCH model	15
3.3	The RS-GARCH model	17
3.4	The GARMS model	18
4	LSW processes and forecasting	20
4.1	LSW processes	20
4.2	Wavelet basis selection	22
4.3	A new forecasting algorithm	23
4.4	An application to volatility forecasting	25
5	Summary of the papers	26
6	Future research	27
6.1	Theoretical development	28
6.2	Empirical applications	29

Acknowledgements

Articles I–V

Appendix: Algorithms and functions

Articles appended to the thesis

The thesis is based on the following articles, which will be refereed to by their Roman numerals.

- I Xie, Y. and Yu, J. 2003. Asymptotics for quasi-maximum likelihood estimators of GARCH(1,2) model under dependent innovations. Research report 2003:5, Centre of Biostochastics, Swedish University of Agricultural Sciences (submitted).
- II Xie, Y. and Yu, J. 2005. Consistency of maximum likelihood estimators for the reduced regime-switching GARCH model. Research report 2005:2, Centre of Biostochastics, Swedish University of Agricultural Sciences (submitted).
- III Xie, Y. 2007. Consistency of maximum likelihood estimators for the regime-switching GARCH model. *Statistics* (to appear).
- IV Xie, Y., Yu, J. and Ranneby, B. 2007. Forecasting using locally stationary wavelet processes. Research report 2007:2, Centre of Biostochastics, Swedish University of Agricultural Sciences (submitted).
- V Xie, Y., Yu, J. and Ranneby, B. 2007. A general autoregressive model with Markov switching: estimation and consistency. Research report 2007:6, Centre of Biostochastics, Swedish University of Agricultural Sciences.

Article III is reproduced by the permission of the publisher.

1 Introduction

Financial time series, i.e. financial outputs collected at successive times, are frequently met in daily life and the scientific world. It is clearly of interest for both practitioners and researchers to study the financial time series, i.e. to understand the mechanism giving rise to the data, and/or to make forecasts of the future values of the series. This thesis is dedicated to statistical inferences of a number of models for financial time series. The main interest in such series is the return series of certain financial instruments, e.g. stock shares, derivatives, share indexes, or foreign exchange rates. There are a number of properties common to such series that are already well known (see, e.g. Cont [18]), including:

- (P1) The mean of the return series is close to zero.
- (P2) The marginal distribution is nearly symmetric, or slightly skewed, with a peak around zero and usually a heavy tail.
- (P3) The autocorrelation of the series itself is insignificant, while the autocorrelation of squares or absolute values of the series is usually significant for a large number of lags.
- (P4) Volatility clustering: it is often observed that big changes (variations) in a series are usually followed by big changes and small changes by small changes.

Almost all these properties can be clearly seen in Figure 1. Figure 1 (top left panel) shows a typical financial return series, the log returns of the daily close Financial Times Stock Exchange 100 Index (FTSE 100) for the London Stock Exchange, from which can be seen volatility clustering. The histogram of this series (top right) reveals that the (empirical marginal) density is almost symmetric, and with a peak around zero. The autocorrelation of the series is insignificant (bottom left), while that of the squared series is significant for at least 35 lags (bottom right). Evidence of heavy tails is more obvious in marginal distributions of individual stock return series than in this index series. For example, the average kurtosis of return series of 367 individual company stock shares included in the Standard & Poor's 500 (S&P500) index from year 1990 to 2001 is around 10.8; some is even as large as 165, compared with the kurtosis for the normal distribution which is 3. This indicates heavy tails for these return series. Properties such as (P3) and (P4) were not handled well by traditional econometric models, until the development of the autoregressive conditionally heteroscedastic (ARCH) model by Engle [25] and the generalized ARCH (GARCH) by Bollerslev [8]. See Section 2 and Cont [18] for more descriptions of the properties of financial time series, and the analysis of these properties using mathematical statistics.

Mathematical statistics is a subject concerned with collecting and gaining information from data, in particular, gaining knowledge about a population by inference from a sample. In practice, such data contain some randomness or uncertainty and mathematical statistics handles this using methods derived from probability theory. For example, statisticians assume that a sample is just one set of the all possible realizations (the ensemble or population) of the actual data generating mechanism, and this mechanism can

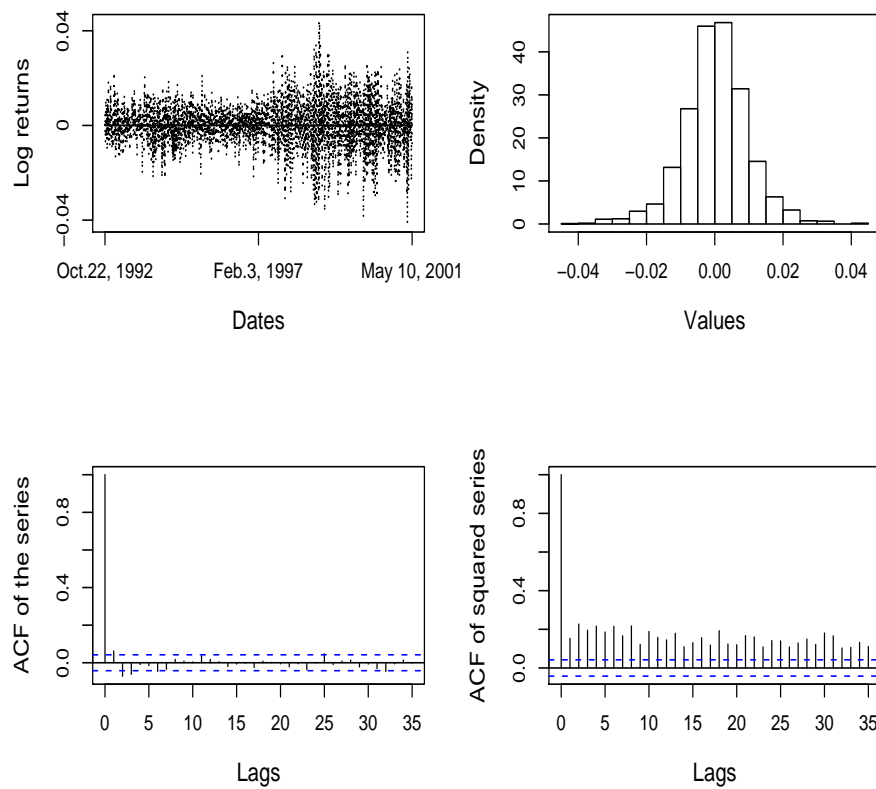


Figure 1: Top left: The plot of the log returns series of the daily close FTSE 100 index with zero line; Top right: The histogram of the series. Bottom left: The autocorrelation function (ACF) of the series; Bottom right: The ACF of squares of the series, where the dashed lines in the bottom figures indicate approximate 95% confidence intervals for the ACF.

be described by a model. Here, a model is a mathematical formulation of a theory that aims to describe the data generating mechanism. In parametric statistical inference, the model usually specifies the probability distribution of the population. But, for this probability distribution, it is assumed that there are several unknown parameters, denoted by θ , the possible values of which are limited within a parameter space Θ . One of the major tasks of statistical inference is to estimate the parameters based on the observed data, so as to more clearly understand the nature of the population, and/or make forecasts on the future events within this population. For more descriptions and other ‘*principles of statistical inference*’, see Cox [19].

There are various methods for estimating the model parameters. The first one that deserves attention is the Least Squares (LS) estimation, also known as Ordinary Least Squares (OLS) estimation. The main idea of the LS estimation, as its name indicates, is to minimize the error impact, which has a quadratic form. In general, suppose we have data $\{\mathbf{x}_i, y_i\}, i = 1, \dots, n$, where \mathbf{x}_i may be a vector. We want to find a function of \mathbf{x}_i with a vectorial parameter β , $f(\mathbf{x}_i, \beta)$, such that $f(\mathbf{x}_i, \beta)$ is “as close as possible” to y_i for all values of i . The LS estimation, intuitively, treats \mathbf{x}_i , y_i , and the form of f as being fixed and estimates the parameter β (in values of \mathbf{x}_i and y_i) such that

$$\sum_{i=1}^n (f(\mathbf{x}_i, \beta) - y_i)^2 \quad (1)$$

is minimized.

It is generally difficult, if not impossible, to obtain an analytical solution for β that minimizes (1) when f is nonlinear and/or some constraints are imposed on β . Instead, we can use numerical optimization methods in such cases. Paper IV presents an example with constrained β that requires a numerical solution. The LS estimation is, however, probably better known in the linear regression analysis. In a regression analysis, the relation between a dependent variable (response variable), Y_i , and certain specified independent variables (explanatory variables), say, X_{i1}, \dots, X_{ip} (p fixed), is to be examined. The relation in a linear regression is assumed to be linear on the parameters β , i.e.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \varepsilon_i, \quad (2)$$

where ε_i is a random error term with mean zero. Many problems may be formulated as model (2). Note that capital letters such as X and Y are usually used to denote random variables and lower case letters denote the values (realizations) of the random variables. If the model (2) is accepted¹, a sample of size n , $\{y_i, x_{i1}, \dots, x_{ip}\}, i = 1, \dots, n$, is obtained, and we want to estimate the parameter $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ (T denotes the transposition of a matrix or vector), such that the sum of square of errors is minimized. Let \mathbf{X} be a $n \times (p+1)$ matrix, the element of which is $x_{ik}, i = 1, \dots, n, k = 0, 1, \dots, p$, with $x_{i0} = 1$ for all i . \mathbf{Y} is the vector $(Y_1, \dots, Y_n)^T$. If the errors ε_i are not correlated to each other, the LS estimation gives us the estimator $\hat{\beta}$ as

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},$$

¹ Note that no model can really be true for real data except in a very general way. It would be wrong here for us to state that ‘the model is true’. It is more appropriate to say ‘the model describes the data well’.

where $^{-1}$ denotes the matrix inverse, provided that this inverse exists.

In this thesis, the focus is on another popular statistical estimation method, Maximum Likelihood (ML) estimation, which has been mainly credited to Sir R. A. Fisher (cf. Aldrich [1]). ML estimation selects the values of model parameters under which the (given) data have a greater probability of being generated than under any other values of the parameters. Suppose that observations $\{y_1, \dots, y_n\}$ are recorded, the probability densities of which depend on some (vectorial) parameter θ . Suppose, in addition, that the joint density of $\{y_1, \dots, y_n\}$ (the likelihood function) is

$$L_n(\theta) = L(y_1, \dots, y_n; \theta). \quad (3)$$

The ML estimator (MLE) is defined as any estimator $\hat{\theta}_n$ that maximizes the likelihood function within some parameter space Θ , i.e.,

$$\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} L_n(\theta).$$

In time series analysis using ML estimation, there can be difficulties when deriving the likelihood function, since observations of time series are usually dependent, and the maximization of likelihood is often complex, see Papers II and III for examples.

In principal, any function of the sample data, including a constant function, can be an estimator for the underlying model parameters. Consequently, an important issue to be considered is the goodness of different estimators. Of the criteria commonly used to evaluate estimators, unbiasedness is easy to understand. It requires that, on average, the estimator, e.g., $\hat{\theta}_n$, as a function of data (random variables) of size n , should be neither bigger nor smaller than the true parameter θ_0 , i.e. the expectation of $\hat{\theta}_n$ (with respect to some measure), $E\hat{\theta}_n = \theta_0$.

Consistency and asymptotic normality of estimators are two other common criteria that concern the asymptotic properties of estimators, i.e. how the estimators behave themselves when more and more (until infinitely many) samples are obtained. A (weakly) consistent estimator tends to the true one in probability when sample size goes to infinity, formally

$$\lim_{n \rightarrow \infty} \operatorname{Pr}(|\hat{\theta}_n - \theta_0| < \epsilon) = 1, \quad (4)$$

for any positive number ϵ . The estimator is said to be strongly consistent if the convergence in probability in the above equation is replaced by convergence with probability one, or almost sure (a.s.) convergence,

$$\operatorname{Pr}\left(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta_0\right) = 1. \quad (5)$$

With asymptotic normality, an estimator is not only consistent, but a clearer picture can be obtained about how quickly (with respect to the number of observations) the estimator will converge to the true parameter; the construction of a confidence interval is also possible. In the usual form, it follows that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Sigma), \text{ as } n \rightarrow \infty, \quad (6)$$

where \xrightarrow{D} denotes convergence in distribution, or, informally, the random variables in the left-hand side of (6) have the same distribution as the right-hand side when n is large enough. Here, $N(\mu, \Sigma)$ is the normal distribution with mean μ and variance Σ .

Another important issue in statistical inferences, in particular for a financial time series, is to make forecasts, i.e. to predict the future values of the series given its past and present values. Suppose that we have observed the values of a series at a number of time points up to and including time T , say, Y_1, Y_2, \dots, Y_T , and want to predict the value of Y_{T+h} ($h > 0$). For example, for an autoregressive (AR) model, it is assumed that Y_t is a linear function of the previous p observations with some random error ε_t that has zero mean and is uncorrelated. The model can be written as $Y_t = \varphi_1 Y_{t-1} + \dots + \varphi_p Y_{t-p} + \varepsilon_t$. Thus, after obtaining the estimates of the parameters, $\hat{\varphi}_i, i = 1, \dots, p$, using certain estimation methods, it is natural to predict the value of Y_{T+1} by $\hat{\varphi}_1 Y_T + \dots + \hat{\varphi}_p Y_{T-p+1}$. This implies that we take the expectation of ε_{T+1} (i.e. zero) in the forecasting, since its value is not available. The forecast of Y_{T+h} can be obtained by substituting the unobserved $Y_{T+i}, i = 1, \dots, h-1$, by their forecasts. See, e.g. Hamilton [39].

It is of interest to know the forecasting accuracy of, say, \hat{Y}_{T+1} as a predictor of Y_{T+1} , which consequently depends on the measure or criterion of the accuracy. The most widely used measure may be the mean square prediction error (MSPE), defined by $E(\hat{Y}_{T+1} - Y_{T+1})^2$. A general result is that a predictor of Y_{T+1} that minimizes the MSPE is the conditional expectation of Y_{T+1} given the previous observations, see, e.g. Priestley [62]. The predictor of an AR model discussed above is a particular example of this general result, when only linear predictors are considered. In practice, to evaluate the usefulness of a forecasting methods or compare the forecasting abilities of different methods, out-of-sample forecasting experiments are usually carried out. In such experiments, while one part of the sample (in-sample) data is used for the model building and parameter estimation, another part (out-of-sample) is deliberately reserved for evaluating the forecasting accuracy. Therefore, the sample MSPEs can be calculated, and different forecasting methods can be compared based on the MSPEs. See Paper IV for examples of forecasting and out-of-sample experiments.

This thesis will mainly address the asymptotic properties of the MLE for certain financial time series models. The results for an ordinary GARCH model with order $(1, 2)$ are obtained under dependent innovations. ML estimations are investigated for two new models: the regime-switching GARCH (RS-GARCH) and reduced RS-GARCH models. In both cases the dependent structures are rather complicated. Consistency of the MLEs is obtained and asymptotic normality is discussed using simulation studies. These results are further extended to a general autoregressive model with regime-switching. The asymptotic results obtained enrich the theory of the GARCH and Markov switching models, and the investigations into ML estimations for the RS-GARCH models facilitate the empirical application of this type of model. In addition, another model, initiated from spectral representations of stationary processes, the Locally Stationary Wavelet (LSW) model is considered. A new forecasting algorithm of LSW processes is proposed. It is applied to real financial data, and shown that it outperforms the existing forecasting algorithm using out-of-sample experiments. Volatility forecasting using this algorithm is also discussed.

These results are discussed in some details in Sections 2, 3, and 4, aiming to achieve a balance between introductory materials and rigorous statistics. A brief summary of the papers and possible future research are presented in Sections 5 and 6. For the reduced RS-GARCH and RS-GARCH models discussed in Section 3, algorithms for data generation, calculation and maximization of the likelihoods were developed. Such a development is not trivial and these algorithms are appended to the thesis. For the LSW modeling, S-plus® codes were available (only) for the Haar wavelet, accompanying the paper by Fryźlewicz et al. [35]; this was valuable for our calculations in Paper IV and is gratefully acknowledged. Algorithms for using other wavelets and the new forecasting algorithm are also appended.

2 Heteroscedastic time series and GARCH models

A time series is a collection of values recorded at sequential time points. For discrete time series, which this thesis focuses on, these time points are often uniformly spaced. Examples of time series include: the daily closing values of stock prices or the index of a stock market, the monthly unemployment reported by authorities, the yearly gross national product (GNP) of a country, as well as the transformations of these data such as log returns of stock shares or GNP growth rates. Random variables in a time series are usually not independent since they are connected by sequential times.

Time series analysis comprises theories and methods for understanding the nature of the underlying random variables, but probably more importantly, the relationships between random variables at different time points and how they develop. For example it is possible to specify a model for a time series, investigate the correlation structure and other properties of the model, discuss estimation of the model's parameters and properties of the estimators, and/or predict future values based on the model and observed values. Besides analysis within the time domain, time series can be analyzed with respect to frequency domain by transforming the data into spectral values using, e.g. Fourier or wavelet transforms.

A basic model in time series analysis, which deserves special emphasis, is the Autoregressive Moving Average (ARMA) model, which models the observation at time t , Y_t , by past observations and moving averages of innovations ε_t as

$$Y_t = \varepsilon_t + \sum_{i=1}^p \varphi_i Y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j},$$

where $\{\varepsilon_t\}$ is usually assumed to be independent and identically distributed (i.i.d.), or a white noise, i.e. ε_t has zero mean and is uncorrelated to ε_s ($s \neq t$), and φ 's and θ 's are constant parameters. The ARMA model has been used intensively and is the benchmark model for time series analysis. For more discussion about the ARMA model and other time series methods, see the classic books by Box and Jenkins [14], Priestley [62], and Brockwell and Davis [16]. For an introduction to the theory of wavelets and their applications to time series analysis, see Daubechies [21] and Percival and Walden [60].

For financial return series from real life, the volatility clustering property (see Figure 1 and (P4) in Section 1) and time varying variation, or heteroscedasticity, have been observed and documented, see, e.g. Fama [28] and McNees [56]. However, traditional time series models cannot explain these properties well and new models are therefore required.

2.1 The GARCH model

For example, suppose Y_t is a financial return series and follows the model

$$Y_t = \gamma + \varepsilon_t, \quad (7)$$

where γ is a constant. In traditional time series analysis, it may be assumed that the innovation $\{\varepsilon_t\}$ is i.i.d. or a white noise. In order to handle the heteroscedasticity, one conventional approach is to assume that this time varying variation comes from another exogenous variable, say, Z_t , i.e.

$$\varepsilon_t = \eta_t Z_t,$$

where $\{\eta_t\}$ is a white noise. This solution is unsatisfactory in that the cause of the time varying variance has to be specified explicitly. Engle [25], instead, proposed the use of previous information and defined the ARCH model with order q as

$$\varepsilon_t = \eta_t h_t^{1/2}, \quad (8)$$

and

$$h_t = \alpha_0 + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2, \quad (9)$$

where η_t has zero mean and unit variance, and the parameters α 's are nonnegative, where α_0 is positive. Like the extension of a moving average (MA) to become the ARMA model, the ARCH model was extended by Bollerslev [8] to become the GARCH model, by allowing the conditional variance h_t further depending on previous conditional variances $\{h_s; s < t\}$. The conditional variance equation of a GARCH(p, q) model is given by

$$h_t = \alpha_0 + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j h_{t-j}, \quad (10)$$

where β 's are also nonnegative. The GARCH model has a more flexible parameter structure than ARCH. In empirical applications, while it is found that a relatively long lag (large q in (9)) is necessary for ARCH models, GARCH(1,1) is usually good enough for describing a large number of financial series, cf. the review by Bollerslev et al. [10]. In one of our experiments, the GARCH characters of daily log return series of stock shares included in S&P500 index were examined. It was found that most series can be modeled by GARCH(1,1), selected by the Akaike information criterion (AIC) among GARCH models, although there are some series that require a more complicated GARCH(1,2) model. The latter in one aspect motivated our study on GARCH(1,2) model in Paper I.

One breakthrough linked to the GARCH model is the association of the observations to the later (conditional) variances, which can be interpreted as returns and risks in financial return series analysis. Therefore, it is no surprise that GARCH-type models quickly gained popularity in the Capital Asset Pricing Model (CAPM), options and derivatives pricing, and other financial fields. At present, GARCH has become the benchmark model for analyzing heteroscedastic time series, see Bollerslev et al. [10] for more information and applications.

2.2 Basic properties of the GARCH model

It is not difficult to check that the GARCH model captures the stylized features of financial time series (P1), (P3), and (P4) presented in Section 1, given that innovations ε_t are uncorrelated. Bollerslev [8] also showed that a GARCH(1,1) process has a heavy tail, i.e. the kurtosis $E(Y_t^4)/E^2(Y_t^2)$ is greater than 3, provided that the fourth moment of ε_t exists and η_t is Gaussian. Under analogous conditions, a general GARCH model has also been shown to have a heavy tail, see, e.g. Fan [29, Proposition 4.2].

For a new time series model, one basic question is whether or not the process is stationary (and ergodic). A weakly stationary process has time-invariant mean and covariance, while strict stationarity requires the joint distribution to be same with any shift in time. Ergodicity is relatively subtler. Recall that usually only one realization (a sample) of a population is observed in statistical inferences, but the properties of the population are to be inferred from the sample. A process is ergodic if and only if the averages over time (over a single realization) converge with probability one to the corresponding ‘ensemble’ averages over many realizations of the process (Priestley [62]). For a stationary Gaussian process $\{Y_t\}_{t \in \mathbb{Z}}$, a sufficient condition for the ergodicity of this process is that its covariance function $Cov(Y_t, Y_{t+h})$ tends to zero, as the lag h goes to infinity.

Bollerslev [8] showed that the necessary and sufficient condition for weak stationarity of the GARCH model ((8) and (10)) is

$$\sum_{i=1}^q \alpha_i + \sum_{j=1}^p \beta_j < 1. \quad (11)$$

For strict stationarity, Nelson [58] found the necessary and sufficient condition for the GARCH(1,1) model to be

$$E(\log(\alpha_1 \eta_t^2 + \beta_1)) < 0. \quad (12)$$

Note that (11) is sufficient for (12), while (12) allows $\alpha_1 + \beta_1$ to be equal to 1 or slightly larger than 1. For the general GARCH model, let $r = \max(p, q)$ and, by convention, the α_i, β_j are equal to zero for $i > q$ and $j > p$. Define

$$\tau_t = (\beta_1 + \alpha_1 \eta_{t-1}^2, \dots, \beta_{r-1} + \alpha_{r-1} \eta_{t-r+1}^2) \in \mathbb{R}^{r-1},$$

and the square matrix A_t of size r in block form as

$$A_t = \begin{pmatrix} \tau_t & \beta_r + \alpha_r \eta_{t-r}^2 \\ I_{r-1} & 0 \end{pmatrix},$$

where I_{r-1} is the identity matrix of size $r - 1$.

Let

$$\mathbf{B}_t = (\alpha_0, 0, \dots, 0)^T \in \mathbb{R}^r$$

and

$$\mathbf{X}_t = (h_t, h_{t-1}, \dots, h_{t-r+1})^T \in \mathbb{R}^r.$$

Then ε_t is a solution of (8) and (10) if and only if \mathbf{X}_t is a solution of

$$\mathbf{X}_t = \mathbf{A}_t \mathbf{X}_{t-1} + \mathbf{B}_t. \quad (13)$$

Bougerol and Picard [12] studied the stationarity conditions for general autoregressive processes in the form (13). They proved that the necessary and sufficient condition for strict stationarity is that the so-called top Lyapunov exponent associated with (\mathbf{A}_t) is negative, or in mathematical notation

$$\rho := \inf_{t \in \mathbb{N}} \left\{ \frac{1}{t+1} E(\log \|\mathbf{A}_0 \cdots \mathbf{A}_{-t}\|) \right\} < 0, \quad (14)$$

where $\|\cdot\|$ is any matrix operator norm. Intuitively, this condition (14) requires that the autoregression of (13) should converge in some sense and the norm of \mathbf{A}_t should be less than 1 “on average”. See also a recent lecture note by Straumann [65] about this result.

Remark 1. In a slightly more general framework, Straumann [65, Proposition 3.3.3] also gave some conditions under which the sign of the Lyapunov exponent ρ can be determined. Among those conditions, a clear conclusion is that the condition for weak stationarity (11) is sufficient for (14) and hence for strict stationarity of the GARCH model. As already seen before, it is true for GARCH(1,1). In fact, Fan [29, Theorem 4.4] showed that condition (11) is not only necessary but also sufficient for a GARCH model to be strictly stationary with finite unconditional variance.

Bollerslev [8] gave the necessary and sufficient conditions for the existence of the even order moments of GARCH(1,1) and for the fourth moments of GARCH(1,2) and GARCH(2,1), provided that ε_t is Gaussian (the odd-order moments are zero). He and Teräsvirta [42] presented a condition for the existence of the fourth moment of a general GARCH model. However, Ling and McAleer [51] pointed out that He and Teräsvirta’s condition was incomplete and derived the necessary and sufficient condition for the existence of all the moments. In general, those conditions are not easy to verify in practice except for small order moments. Hence they are not cited here and only referred to the aforementioned works and the review by Li et al. [49].

2.3 Quasi-MLE

The estimation of GARCH models is usually carried out using ML estimation. However, obtaining a handleable likelihood function is not straightforward. The following simple fact,

$$p(X_1, \dots, X_n) = p(X_n | X_{n-1}, \dots, X_1) \cdot p(X_{n-1} | X_{n-2}, \dots, X_1) \cdots p(X_1), \quad (15)$$

often helps to determine the joint density of dependent random variables in time series analysis, where $p(\cdot)$ is used to denote joint, marginal, or conditional density according to the context, assuming this does not cause confusion.

The distribution of η_t has also to be specified to derive a likelihood. Note that even when $\{\eta_t\}$ is assumed to be i.i.d. and Gaussian, the distribution of ε_t is still unknown since the distribution of h_t is not known. A common practice in estimation of the GARCH models is to assume η_t to be Gaussian when deriving the likelihood, making use of (15), and then to discard this assumption later. This method is called Quasi Maximum Likelihood (QML) estimation and is now a basic estimation method for classic GARCH models. To the author's knowledge, the idea is related to the work by Wedderburn [71], in which Wedderburn defined a Quasi Likelihood function for linear models, based only on the mean and the variance structure of observations. He proved that this quasi likelihood had properties similar to the log likelihood and they were the same for a one-parameter exponential family. Another common practice in QML estimation for GARCH models is just to obtain the (conditional) mean and variance of observations and then insert them into a Gaussian density. It is worth noting that in a slightly different context, there is a related method called Quasi-likelihood estimating function. It aims to find an optimal (with some criteria) estimating function (instead of the estimates of parameters) for diversified models under predetermined functional space. Readers who are interested in this topic are referred to the monograph by Heyde [43].

To the GARCH model (8) and (10), apply the hypothesis that $\{\eta_t\}$ is i.i.d. $N(0, 1)$. Under this synthetic assumption, $\varepsilon_t | \varepsilon_{t-1}, \dots$ is conditionally normally distributed with zero mean and conditional variance h_t . Suppose now we are given observations $\{y_1, \dots, y_n\}$. Observe that for some conditional densities, for instance, $p(\varepsilon_1)$, the initial observations, i.e. y_0, \dots, y_{1-q} are unavailable and h_0, \dots, h_{1-p} are unobservable. The likelihood has to be conditional on these initial values, which in practice can be set to, say, zero for y_0, \dots, y_{1-q} and $\alpha_0/(1 - (\beta_1 + \dots + \beta_p))$ for h_0, \dots, h_{1-p} (recall that $\beta_1 + \dots + \beta_p < 1$ is necessary for strict stationarity of the GARCH model according to Bougerol and Picard [13]). Straumann [65] showed that these initial values would not affect the asymptotic properties of the ML estimation, see also Straumann and Mikosch [66].

Following on from the above points, use (15) and the Gaussian hypothesis of η_t . The log QML function of the GARCH model ((7)-(8) and (10)) is (ignoring some constant) a function of θ (with y_1, \dots, y_n fixed),

$$L_n(\theta) = L(y_1, \dots, y_n; \theta) = -\frac{1}{2} \sum_{t=1}^n \left(\frac{(y_t - \gamma)^2}{h_t} + \log h_t \right), \quad (16)$$

where the parameter $\theta = \{\gamma, \alpha_0, \alpha_1, \dots, \alpha_q, \beta_1, \dots, \beta_p\}$. The QML estimator (QMLE) $\hat{\theta}_n$ is defined as any maximizer of $L_n(\theta)$ within some parameter space Θ , which will be specified later. Note that in many cases, the GARCH models focus only on the variance structure and $\{\varepsilon_t\}$ are assumed to be the observations, or equivalently $\gamma = 0$ in (7); this changes the parameters slightly.

2.4 Asymptotic properties

The asymptotic properties of QMLEs for the ARCH and GARCH(1,1) models have been studied by, amongst others, Weiss [72] and Lumsdaine [53], respectively. The results for the general GARCH(p, q) model were completed by the work of Berkes et al. [5] and

Francq and Zakoian [31]. See the review by Li et al. [49] and references therein for more results.

2.4.1 Asymptotics for i.i.d. η_t

Assume $\{\eta_t\}$ is i.i.d. and $\gamma = 0$. The parameter space Θ_1 is a compact subset of $0 \times (0, \infty) \times [0, \infty)^p \times B$, where $B := \{(\beta_1, \dots, \beta_p)^T \in [0, 1)^p \mid \sum_{j=1}^p \beta_j < 1\}$. The QMLE $\hat{\theta}_n$ maximizes the likelihood (16) under Θ_1 . Write two polynomials $\mathcal{A}_\theta(z) = \sum_{i=1}^q \alpha_i z^i$ and $\mathcal{B}_\theta(z) = 1 - \sum_{j=1}^p \beta_j z^j$. The typical assumptions include

- (C1) η_t^2 has a non-degenerate distribution with $E\eta_t^2 = 1$.
- (C2) The true parameter $\theta_0 \in \Theta_1$.
- (C3) Under θ_0 , the top Lyapunov exponent ρ defined in (14) is strictly negative.
- (C4) The polynomials $\mathcal{A}_{\theta_0}(z)$ and $\mathcal{B}_{\theta_0}(z)$ have no common root. $\mathcal{A}_{\theta_0}(1) \neq 0$ and the true values of α_q and β_p are not zero.
- (C5) θ_0 is in the interior of Θ_1 .
- (C6) $\kappa_\eta := E\eta_t^4 < \infty$.

Assumption (C4) ensures the identifiability of the model parameters, while others are self-explanatory. A standard asymptotic result (Theorem 2.4.1) for GARCH models follows. It is probably the result that requires the ‘weakest’ assumptions (cf. Li et al. [49]), while Berkes et al. [5] needed the $(2 + \delta)$ -th ($\delta > 0$) moment condition for consistency and $(4 + \delta)$ -th for asymptotic normality.

Theorem 2.4.1 (Francq and Zakoian [31, Theorems 2.1 and 2.2]) *Let $(\hat{\theta}_n)$ be a sequence of QMLEs. Under Assumptions (C1)-(C4), $\hat{\theta}_n \rightarrow \theta_0$ a.s., as $n \rightarrow \infty$. If in addition (C5) and (C6) are satisfied, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Sigma),$$

as $n \rightarrow \infty$, where

$$\Sigma = (\kappa_\eta - 1) E_{\theta_0}^{-1} \left(\frac{1}{h_t^2(\theta_0)} \frac{\partial h_t^2(\theta_0)}{\partial \theta} \frac{\partial h_t^2(\theta_0)}{\partial \theta^T} \right).$$

2.4.2 Asymptotics for dependent η_t

Since only the mean and variance structure are needed and correctly specified for the QML estimation of GARCH models, the independence assumption of η_t can be relaxed. Lee and Hansen [47], therefore, investigated the asymptotic properties for the GARCH(1,1) model under stationary (and ergodic) innovations. Xie and Yu (Paper I) extended these results to the GARCH(1,2) model under weaker conditions. Define $\hat{\theta}_n$ as any maximizer of (16) for the GARCH(1,2) model (7) and (10) under

$$\Theta_2 = \{ \theta : \gamma_l \leq \gamma \leq \gamma_u, 0 < \alpha_{0l} \leq \alpha_0 \leq \alpha_{0u}, 0 < \alpha_{1l} \leq \alpha_1 \leq \alpha_{1u}, \\ 0 < \alpha_{2l} \leq \alpha_2 \leq \alpha_{2u}, 0 < \beta_l \leq \beta \leq \beta_u < 1 \},$$

where $\gamma_l, \gamma_u, \alpha_{0l}, \alpha_{0u}, \alpha_{1l}, \alpha_{1u}, \alpha_{2l}, \alpha_{2u}, \beta_l$ and β_u are constants. Note that Θ_2 is not necessarily smaller than Θ_1 . Assume the true parameter $\theta_0 \in \Theta_2$.

Theorem 2.4.2 (Xie and Yu [Paper I, Theorems 1 and 2]) *Consider $p = 1$ and $q = 2$ in the GARCH model (10) and η_t is strictly stationary and ergodic. Besides Assumption (C1), assume that the true parameters of α_1, α_2 and β_1 satisfy*

$$\alpha_{10} + \alpha_{20} + \beta_{10} < 1.$$

Then $\hat{\theta}_n \rightarrow \theta_0$ a.s., as $n \rightarrow \infty$. If we further assume that θ_0 is in the interior of Θ_2 and

$$E(\eta_t^4 | \mathcal{F}_{t-1}) < \kappa < \infty,$$

then $\hat{\theta}_n$ is asymptotically normal, where \mathcal{F}_{t-1} denotes the information up to time $t - 1$ and κ is a positive constant.

Remark 2. Lee and Hansen [47] assumed

$$\sup_t E(\log(\beta_{10} + \alpha_{10}\eta_t^2) | \mathcal{F}_{t-1}) < 0 \text{ a.s.}$$

and a uniformly finite conditional $(2+\delta)$ -th moment for a local consistency result (under a subset of Θ_2) for the QMLE of the GARCH(1,1) model. Their global consistency was also based on (*inter alia*) $\alpha_{10} + \beta_{10} < 1$. Note that we need only the second moment condition for the consistency. In this sense, our assumptions are weaker than theirs. Note that the condition $\alpha_{10} + \beta_{10} < 1$ is sufficient for $\sup_t E(\log(\beta_{10} + \alpha_{10}\eta_t^2) | \mathcal{F}_{t-1}) < 0$ a.s. See Remark 1 for a discussion about these two conditions.

2.5 Other models

There are many extensions to the standard ARCH and GARCH models (8)-(10). Some are mainly of theoretical interest, for example, the continuous time GARCH processes of Klüppelberg et al. [45] and Brockwell et al. [15], while most of them are more for practical (economical and/or financial) considerations. For example, it is believed in finance that bad news have bigger impacts to the volatility of financial time series than good news. This so-called Leverage Effect has motivated the development of the asymmetric GARCH extensions, including the Exponential GARCH (EGARCH) of Nelson [59], GJR-GARCH developed by Glosten et al. [36], Quadratic GARCH developed by Sentana [63], Threshold GARCH (TGARCH) developed by Zakoian [75] and others, and Asymmetric GARCH developed by Ding et al. [23] and Straumann [65] amongst others. The consideration of portfolio management leads to multivariate GARCH models and multivariate extensions of other GARCH extensions, see (*inter alia*) Bollerslev et al. [11] and Bollerslev [9]. Another frequently observed feature in applied work is the volatility persistence, implied by the estimates of parameters in the variance equation, where the estimate of $\sum_{i=1}^q \alpha_i + \sum_{j=1}^p \beta_j$ is close to 1. The strong persistence underlying the GARCH model can be possibly explained by the Integrated GARCH (IGARCH) by Engle and Bollerslev [26], Fractional IGARCH (Baillie [3]), and the RS-GARCH models (Hamilton [38], our Papers II and III) that are introduced in Section 3.

Not only to the conditional variance equation (10), there are also extensions made to the mean structure. Sometimes, it seems restrictive to assume that the observed process

is a pure GARCH. Therefore, it is natural to view the GARCH as an error process, as in the original works of Engle [25] and Bollerslev [8]. It is also possible to use the GARCH as an innovation to an ARMA model. This extension entails essential difficulties. Some theoretical results can be found in Ling and Li [50], Ling and McAleer [52], and Francq and Zakoian [31]. Another similar extension is the so-called ARCH-in-Mean model, where the conditional variance h_t is directly entered into the regression equation (7), see Engle et al. [27]. A full description of such extensions is too lengthy for this thesis, so readers are referred to Bollerslev et al. [10] and Teräsvirta [70] and references therein for applications of GARCH models and the extensions.

Finally, it may be worth mentioning an alternative model to GARCH, namely the Stochastic Volatility (SV) model. A simple SV model, from Taylor [69], is defined as

$$\varepsilon_t = h_t \eta_t$$

and

$$\log h_t = \alpha + \beta \log h_{t-1} + \xi_t, \quad (17)$$

where η_t is independent of h_t . It can be seen that the SV model allows more flexible parameters (compared with the rather restrictive GARCH). However, there is a random term (ξ_t) entangled with the already unobservable conditional variance h_t , and the estimation of the SV model is complicated. We do not pursue the SV model further, but refer readers to a review by Shephard [64] on SV and ARCH models.

3 Markov switching models

Nowadays, an increasing number of economic researchers think it may be more reasonable to consider that in the long term there are different economic states, and the outcomes of an economic system depend on these states. For example, in the analysis of the US annual GNP growth rate series, Hamilton [38] treated the expansion and recession period as two states (called regimes in econometric literature) and proposed a model in the form

$$Y_t - \mu(R_t) = \sum_{i=1}^s \beta_i (Y_{t-i} - \mu(R_{t-i})) + \varepsilon_t,$$

where $\{Y_t\}$ are the observations in question, β_i , $i = 1, \dots, s$, are coefficients, R_t is the regime at time t , $\mu(R_t)$ are constants depending on the regimes R_t (assuming two regimes in his model) and ε_t is distributed as $N(0, \sigma^2)$. Hamilton assumed that the regimes are unobservable and the shift between the two regimes is governed by a Markov chain (unobserved). The inference of the model has to be based only on the observations, the outcomes of some economic variables. Hamilton [38] showed that such Markov switching or regime switching models have advantages in model interpretation and data description. The idea of Markov switching has drawn much attention during the last two decades from both statisticians and practitioners: see the review by Hamilton and Raj [40].

It is worth noting that Markov switching models are closely related to the Hidden Markov Model (HMM), which has been very popular in fields such as engineering,

biology and statistics. Unlike the Markov switching model in which Y_t is dependent on Y_{t-1} , the observations in a HMM are independent given the corresponding regimes. Hence, HMM has a simpler dependent structure than the Markov switching model. See the monograph by MacDonald and Zucchini [54] for a comprehensive introduction to HMM.

3.1 Regime-switching GARCH and path dependence problem

Some attempts (e.g. Cai [17] and Hamilton and Susmel [41]) have been made to incorporate the Markov Switching into the popular GARCH models. Keeping our attention only on the conditional variance structure, the regime-switching GARCH (RS-GARCH) process $\{Y_t\}_{t \in \mathbb{Z}}$ is defined by

$$\begin{aligned} Y_t &= (h_t)^{1/2} \eta_t, \\ h_t &= \omega(R_t) + \sum_{i=1}^q \alpha_i(R_t) Y_{t-i}^2 + \sum_{j=1}^p \beta_j(R_t) h_{t-j}, \end{aligned} \quad (18)$$

where $\{\eta_t\}$ as usual is a sequence of i.i.d. random variables with zero mean and unit variance. $\{R_t\}$ is a Markov chain with a finite state space $\mathbf{E} = \{1, 2, \dots, d\}$. Given $\{R_t = s\}$, $s \in \mathbf{E}$, Y_t follows a GARCH model. As in an ordinary GARCH model, assume parameters $\alpha_i(s) \geq 0$, $\beta_j(s) \geq 0$, and $\omega(s) > 0$, for all i, j, s . Suppose that $\{\eta_t\}$ and $\{R_t\}$ are independent and that the Markov chain is irreducible and aperiodic with stationary distribution $\pi(s) := P(R_1 = s)$, $s \in \mathbf{E}$, and transition probabilities $p_{kl} := P(R_t = l | R_{t-1} = k)$. Also note that $\pi(s) > 0$ for all $s \in \mathbf{E}$ under the assumptions. The Markov chain $\{R_t\}$ is unobservable and its transition probabilities and stationary distribution are unknown. Our aim is to draw statistical inference based only on the observed $\{Y_t\}$.

ML estimation may be used for the estimation of the RS-GARCH model. Suppose that we are given a realization $\{y_1, \dots, y_n\}$. Making use of (15) and taking into account the regime switching, summing up the (conditional) probability density over all possible paths of the Markov chain leads to a likelihood function (r_t denotes the value of R_t)

$$L_n(y_1, \dots, y_n; \theta) = \sum_{(r_1, \dots, r_n) \in \mathbf{E}^n} \pi(r_1) \left\{ \prod_{t=2}^n p_{r_{t-1}, r_t} \right\} \left\{ \prod_{t=1}^n f_{r_1, \dots, r_t}(y_1, \dots, y_t) \right\}, \quad (19)$$

where $f_{r_1, \dots, r_t}(y_1, \dots, y_t)$ is the conditional density of Y_t given previous observations and regimes and

$$\theta := \{p_{kl}, \omega(s), \alpha_i(s), \beta_j(s), k \neq l, 1 \leq k, l, s \leq d, 1 \leq i \leq q, 1 \leq j \leq p\},$$

which contains the transition probabilities of the Markov chain and parameters of the GARCH equation (18). The stationary distribution $\pi(s)$, $s \in \mathbf{E}$, is not included since, asymptotically, the stationary distribution will not affect the estimation (see e.g. Leroux [48]). We assume that p, q and d are known.

However, the likelihood (19) is not easy to handle since the number of possible regime paths grows exponentially with t . The likelihood becomes intractable very quickly as n increases. This regime path dependence problem is induced by the the

recursive structure in (18): h_t depends on the whole regime path through h_{t-j} and further on. The applicability of the RS-GARCH model is limited. Hamilton and Susmel [41], Cai [17], and Francq et al. [32] had to limit their estimation to the RS-ARCH model, i.e. letting $p = 0$ in (18). Two possible alternatives will be discussed in Sections 3.2 and 3.3.

3.2 The reduced RS-GARCH model

In order to overcome the path dependence problem, Gray [37] proposed a reduced model for RS-GARCH(1,1), where (18) is replaced by

$$h_t = \omega(R_t) + \alpha(R_t)Y_{t-1}^2 + E_{\tilde{R}_{t-1}}[\beta(R_t)h_{t-1}],$$

and the expectation is across the regime path $\tilde{R}_{t-1} := \{R_{t-1}, R_{t-2}, \dots\}$, conditional on available information up to time $t-1$, \mathcal{F}_{t-1} . Note that Y_t (Δr_t in Gray's context) is essentially a mixture of distributions with respect to different regimes (with time-varying mixing parameters), it is natural to consider to take expectation of individual conditional variances over regimes. This integrated variance is then used as the lagged conditional variance in constructing the conditional variance of the next time period, and the path dependence problem can be overcome while the essential nature of the GARCH process is preserved.

Gray's idea is generalized to GARCH(p, q) model by Xie and Yu [Paper II]. It is referred to as the reduced RS-GARCH, in which (18) is replaced by

$$Y_t = (h_t)^{1/2}\eta_t,$$

$$h_t = \omega(R_t) + \sum_{i=1}^q \alpha_i(R_t)Y_{t-i}^2 + E_{\tilde{R}_{t-1}}\left[\sum_{j=1}^p \beta_j(R_t)h_{t-j}\right]. \quad (20)$$

Note that actually we only need to integrate out the single regime R_{t-1} at time point t since recursively h_{t-1} is already independent of \tilde{R}_{t-2} . For the reduced RS-GARCH model, the conditional density f in (19) depends only on the current regime. Because of the simplified structure, the likelihood for this model can be written as a product of matrices as

$$L_n(y_1, \dots, y_n; \theta) = \mathbf{1}^T \left\{ \prod_{t=2}^n \mathbf{M}_\theta(y_1, \dots, y_t) \right\} \mathbf{p}, \quad (21)$$

where $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^d$, $\mathbf{p} = (\pi(1)f_1(y_1), \dots, \pi(d)f_d(y_1))^T \in \mathbb{R}^d$ and matrix

$$\mathbf{M}_\theta(y_1, \dots, y_t) = \begin{pmatrix} p_{11}f_1(y_1, \dots, y_t) & p_{21}f_1(y_1, \dots, y_t) & \cdots & p_{d1}f_1(y_1, \dots, y_t) \\ p_{12}f_2(y_1, \dots, y_t) & p_{22}f_2(y_1, \dots, y_t) & \cdots & p_{d2}f_2(y_1, \dots, y_t) \\ \cdots & \cdots & \cdots & \cdots \\ p_{1d}f_d(y_1, \dots, y_t) & p_{2d}f_d(y_1, \dots, y_t) & \cdots & p_{dd}f_d(y_1, \dots, y_t) \end{pmatrix}.$$

Assuming a Gaussian innovation process η_t , the MLE $\hat{\theta}_n$ is defined as any maximizer of (21) within parameter space Θ_3 , which is a compact subset of some Euclidean space. The likelihood (21) can be calculated using, e.g. the so-called forward-backward

algorithm. It turns out that it is important to compute the conditional probabilities $\lambda_{st} := P(R_t = s | \mathcal{F}_{t-1})$, $s \in \mathbf{E}$. The integrated variance at time t is then the weighted average of h_t over different regimes s with respect to weight λ_{st} , and those integrated variances will be used in the conditional density and in the right-hand side of (20) as the lagged variances. See Xie and Yu [Paper II] for the recursive formula of λ_{st} and the Appendix for the algorithm to calculate the likelihood (21).

The consistency of MLE for the reduced RS-GARCH model is obtained by Xie and Yu [Paper II] under the assumptions

- (D1) The true parameter $\theta_0 \in \Theta_3$
- (D2) $\{Y_t\}_{t \in \mathbb{Z}}$ in model (20) is strictly stationary and ergodic. In addition, the unconditional variance of Y_t is finite.
- (D3) For any θ_1 and $\theta_2 \in \Theta_3$ and all Y_t, Y_{t-1}, \dots , if $p(Y_t | Y_{t-1}, Y_{t-2}, \dots; \theta_1) = p(Y_t | Y_{t-1}, Y_{t-2}, \dots; \theta_2)$, a.s. under the true parameter θ_0 , then $\theta_1 = \theta_2$.

Assumption (D3) ensures the identifiability of parameters in (20). The following theorem generalizes the results for HMM by Leroux [48].

Theorem 3.2.1 (Xie and Yu [Paper II, Theorem 1]) For the reduced RS-GARCH model (20), assume (D1)-(D3). Then

$$\hat{\theta}_n \rightarrow \theta_0, \text{ a.s. as } n \rightarrow \infty.$$

The asymptotic normality of the MLE is not proved in Xie and Yu [Paper II], but in their simulation study, the Quantile-Quantile (QQ) plot (Figure 2) and the Kolmogorov-Smirnov goodness-of-fit test suggest that the MLE for a reduced RS-GARCH(1,1) model is asymptotically normally distributed. See Xie and Yu [Paper II] for details.

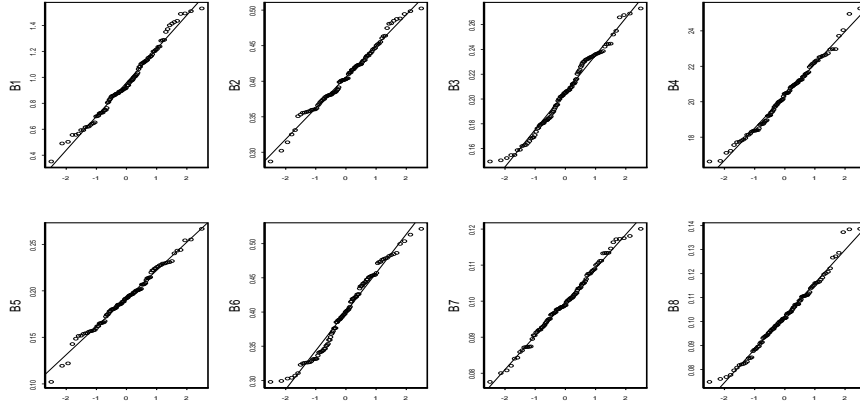


Figure 2: The QQ-plot of MLE for a reduced two-regime switching GARCH(1, 1) model with 0-1 line: B1–B8 represent estimators of $\omega(1)$, $\omega(2)$, $\alpha_1(1)$, $\beta_1(1)$, $\alpha_1(2)$, $\beta_1(2)$, p_{12} and p_{21} , respectively. See Xie and Yu [Paper II] for more information.

3.3 The RS-GARCH model

The regime path dependence problem of the RS-GARCH model mentioned in Section 3.1 is derived from the GARCH-part of the conditional variance equation (18). For ARMA, it is well known that an invertible ARMA process can be transformed into an AR process with infinite order. See Brockwell and Davis [16] for a description of the property of invertibility and this result. The similarity between GARCH and ARMA models (cf. Bollerslev [8]) indicates a possible analogous treatment of the RS-GARCH model.

As with the discussion of the ordinary GARCH in Section 2.2, the condition under which Y_t is stationary for the RS-GARCH model can be investigated by writing (18) in the form of (13). Let \mathbf{A}_t^* , τ_t^* , and \mathbf{X}_t^* be defined as in Section 2.2 except that now they depend on the current regime R_t . Define $\mathbf{B}_t^* = (\omega(R_t), 0, \dots, 0)^T \in \mathbb{R}^r$. Then Y_t is a solution of (18) if and only if \mathbf{X}_t^* is a solution of

$$\mathbf{X}_t^* = \mathbf{A}_t^* \mathbf{X}_{t-1}^* + \mathbf{B}_t^*(R_t). \quad (22)$$

Mimicking the argument of Bougerol and Picard [12], Francq et al. [32, Theorem 1] proved that (22) has a strictly stationary solution if and only if the top Lyapunov exponent ρ^* associated with (\mathbf{A}_t^*) is negative, i.e.

$$\rho^* = \inf_{t \in \mathbb{N}} \left\{ \frac{1}{t+1} E(\log \|\mathbf{A}_0^* \cdots \mathbf{A}_{-t}^*\|) \right\} < 0. \quad (23)$$

Assume that

(E1) The random variable η_t is non-degenerated.

(E2) $\sup_{\theta \in \Theta_4} E_{\theta_0} [|\log p_{\theta}(Y_1)|] < \infty$,

where θ is the set of model parameters as in the reduced RS-GARCH model (Section 3.2), the parameter space Θ_4 is a compact subset of some Euclidean space, and the true parameter $\theta_0 \in \Theta_4$. Under Assumptions (E1) and (E2), Berkes et al. [5] showed that for a strictly stationary and ergodic solution of (18), we have

$$h_t = c_0(R_t) + \sum_{i=1}^{\infty} c_i(R_t) Y_{t-i}^2, \quad \forall t \in \mathbb{Z}$$

with probability one and this representation is unique with coefficients

$$c_0(R_t) = \frac{\omega(R_t)}{\mathcal{B}_t^*(1)}$$

and

$$c_n(R_t) = \frac{d^n}{dx^n} \left(\frac{\mathcal{A}_t^*(x)}{\mathcal{B}_t^*(x)} \right)_{x=0}, \quad 1 \leq n < \infty,$$

where $\mathcal{A}_t^*(x) = \sum_{i=1}^q \alpha_i(R_t) x^i$ and $\mathcal{B}_t^*(x) = 1 - \sum_{j=1}^p \beta_j(R_t) x^j$.

By transforming the RS-GARCH into an (infinite order) RS-ARCH model, a handleable likelihood as (21) can be obtained. The ML estimation is hence possible. Recursive formulae for $\{c_i\}$ are available in Berkes et al. [5], while the algorithms for the transformation and calculation of the likelihood are presented in the Appendix of this thesis. Under Gaussian innovations η_t and Assumption (E2) (*inter alia*), Xie [Paper III] proved the consistency of the MLE. Xie [Paper III] also conjectured the consistency for non-Gaussian innovations and provided numerical evidence by using two-component mixture normal distributions for η_t . The density is

$$(1 - p)\phi(\mu_1, \sigma_1^2) + p\phi(\mu_2, \sigma_2^2).$$

A rather unusual example is reported with the true parameters $\mu_1 = 0.25, \sigma_1^2 = 0.3, \mu_2 = -2.25, \sigma_2^2 = 1.675$, and $p = 0.1$, which implies a zero mean, unit variance, skewness around -2.05 , kurtosis 8.84 , and two modes. From Table 1, it follows that while the biases usually decrease, the standard deviations always decrease as the sample size increases, which suggests that the estimates are consistent.

Table 1: The standard deviation and bias (in parentheses) of MLE for the regime-switching GARCH model with mixture normal innovations, for sample sizes $n = 500$, 2000 , and 5000 , respectively.

n	$\hat{\omega}(1)$	$\hat{\alpha}_1(1)$	$\hat{\beta}_1(1)$	$\hat{\omega}(2)$	$\hat{\alpha}_1(2)$	$\hat{\beta}_1(2)$	\hat{p}_{12}	\hat{p}_{21}
500	0.315 (0.011)	0.230 (0.064)	0.127 (0.031)	12.24 (6.83)	0.186 (0.138)	0.225 (-0.012)	0.042 (0.014)	0.037 (-0.005)
2000	0.148 (-0.01)	0.131 (0.01)	0.074 (0.042)	6.839 (5.538)	0.126 (0.155)	0.134 (0.004)	0.015 (0.016)	0.014 (-0.001)
5000	0.091 (-0.004)	0.082 (0.008)	0.045 (0.035)	4.984 (4.849)	0.064 (0.132)	0.105 (0.016)	0.011 (0.013)	0.012 (-0.004)

As in the case of the reduced RS-GARCH case, the conjecture of the asymptotic normality of the MLE is also supported, in view of the QQ-plot and goodness-of-fit tests, see Xie [Paper III]. Bickel et al. [6] proved asymptotic normality for the general HMM, but the conditional independence of observations given the unobserved regime is crucial in their framework and cannot be easily relaxed to include our model. Douc et al. [24] obtained the asymptotic normality for a class of autoregressive models with Markov regime, which includes HMM and the regime-switching ARCH model as special cases. However, they use the Markov property of $\{R_k, Y_k, \dots, Y_{k-s+1}\}$ (assume an s -th order ARCH model), which does not hold for GARCH or infinite order ARCH models. A proof of the asymptotic normality of the MLE is still needed.

3.4 The GARMS model

A closer examination of the consistency result for the reduced RS-GARCH (Xie and Yu [Paper II]) and RS-GARCH (Xie [Paper III]) seems to indicate that not only the Gaussian assumption for innovations, but also the particular model structure are not essential for the proof. It suggests that it should be possible to extend those results to the following general autoregressive model with Markov switching (GARMS) defined by

$$Y_t = f_\theta(\bar{\mathbf{Y}}_{t-1}^s, R_t; \varepsilon_t), \quad (24)$$

where $\{\varepsilon_t\}$ is an i.i.d. innovation process, \bar{Y}_{t-1}^s denotes observations $\{Y_s, \dots, Y_{t-1}\}$ from possibly an infinite past, i.e. s may be equal to $-\infty$. By convention, this set is empty when $s > t-1$. $\{R_t\}_{t \in \mathbb{Z}}$ is an unobservable Markov chain with finite state space $\mathbf{E} = \{1, 2, \dots, d\}$, where d is known and fixed. Its transition probability matrix is $\mathbf{A} = (p_{kl})$, where $p_{kl} = P(R_t = l | R_{t-1} = k)$, $k, l \in \mathbf{E}$. The parameter θ may depend on the regimes R_t , and is assumed to be finite-dimensional. An infinite-dimensional setting sounds attractive but is technically formidable. f_θ is a family of measurable functions indexed by θ and has implicit requirements imposed by the (conditional) density of Y_t . One of the examples of (24) is the infinite order AR model with Markov switching, see Xie et al. [Paper V].

Clearly, when $s > t-1$, i.e. the conditional distribution of Y_t does not depend on lagged Y 's but only on R_t , model (24) leads to the aforementioned HMM. Yao and Attali [74] studied the stability of this process when $s = t-1$, i.e. a first order autoregression in (24), including conditions under which there is a stationary solution and finite moments of $\{Y_t\}$ and for which limit theorems can be applied. Francq and Rousignol [30] also considered the stability of the process and the consistency of the MLE in this case. A natural and interesting case is that of s less than $t-1$ but finite. The finite order autoregressive model with Markov switching is called ARMS model. Krishnamurthy and Rydén [46] obtained the consistency for the MLE of the ARMS model when the Markov chain of this model has finite states. Douc et al. [24] not only extended it to continuous state space but also proved the asymptotic normality of MLE for both stationary and non-stationary observation sequences. Note that the ARMS model also includes the finite order RS-ARCH model, studied by, among others, Cai [17] and Hamilton and Susmel [41].

Given the distribution of ε_t , the regime $R_t = k$, observations \bar{Y}_{t-1}^s and function f_{θ_k} , assume that the conditional distribution of Y_t has a density $q(y_t | \bar{Y}_{t-1}^s; \theta_k)$ with respect to some Lebesgue measure. Here $\theta_k, k \in \mathbf{E}$, belong to some finite-dimensional parameter space Θ . Denote the whole model parameter including those of the Markov chain $\{R_t\}$ as ϕ , which belongs to Φ , a subset of some Euclidean space. That is, formally we have $\mathbf{A}(\phi) = (p_{kl}(\phi))$ and $\theta_k(\phi) \in \Phi$ for $k, l \in \mathbf{E}$. The usual case is just $\phi = \{p_{11}, p_{12}, \dots, p_{dd}, \theta_1, \dots, \theta_d\}$ (θ_k may be a vector), and $p_{kl}(\cdot)$ and $\theta_k(\cdot)$ equal to coordinate projections. The true parameter is denoted by ϕ_0 and assume $\phi_0 \in \Phi$.

Define $p(Y_t | \bar{Y}_{t-1}^{-\infty}; \phi)$ as the conditional density of Y_t given $\bar{Y}_{t-1}^{-\infty}$ under ϕ and its logarithm as $g(Y_t | \bar{Y}_{t-1}^{-\infty}; \phi)$. The technical assumptions include

- (A1) The Markov chain $\{R_t\}_{t \in \mathbb{Z}}$ is aperiodic and irreducible.
- (A2) $\{Y_t\}_{t \in \mathbb{Z}}$ is a strictly stationary and ergodic process.
- (A3) For each k and $l \in \mathbf{E}$, $p_{kl}(\cdot)$ and $\theta_k(\cdot)$ are continuous on Φ , and $q(y_t | \bar{Y}_{t-1}^s; \theta_k(\phi))$ is continuous on Φ for all realizations of \bar{Y}_{t-1}^s .
- (A4) For any $\phi \in \Phi$, $0 < \min_{k \in \mathbf{E}} q(Y_t | \bar{Y}_{t-1}^{-\infty}; \theta_k(\phi)) \leq \max_{k \in \mathbf{E}} q(Y_t | \bar{Y}_{t-1}^{-\infty}; \theta_k(\phi)) < \infty$ for all $\bar{Y}_{t-1}^{-\infty}$ a.s. under ϕ_0 ; and there exists a neighborhood of ϕ , $V(\phi) = \{\phi' : d(\phi, \phi') \leq \delta\}$ for some $\delta > 0$ and the Euclidean distance $d(\cdot, \cdot)$, such that $E_{\phi_0} \left[\sup_{\phi' \in V(\phi)} |g(Y_t | \bar{Y}_{t-1}^{-\infty}; \phi')| \right] < \infty$.

(A5) Identifiability condition: For any ϕ_1 and $\phi_2 \in \Phi$, if for all $\bar{\mathbf{Y}}_t^{-\infty}$, $p(Y_t | \bar{\mathbf{Y}}_{t-1}^{-\infty}; \phi_1) = p(Y_t | \bar{\mathbf{Y}}_{t-1}^{-\infty}; \phi_2)$ a.s. under ϕ_0 , then $\phi_1 = \phi_2$.

These assumptions are fairly standard in the context of ARMS. They are also found, for example, in Francq and Roussignol [30], Krishnamurthy and Rydén [46], and Douc et al. [24].

Once again, ML estimation is utilized. The likelihood has the form

$$L_n^*(y_n, \dots, y_1; \phi) = \sum_{(r_1, \dots, r_n)} \pi(r_1) \left\{ \prod_{t=2}^n p_{r_{t-1}, r_t} \right\} \left\{ \prod_{t=1}^n q(y_t | \bar{\mathbf{Y}}_{t-1}^1; \theta_{r_t}(\phi)) \right\}. \quad (25)$$

The MLE is defined as any parameter $\hat{\phi}_n$ that maximizes the likelihood L_n^* over a compact subset of Φ , Φ^* . The consistency of the MLE is obtained in the following theorem. Simulation studies using a particular infinite order AR model with Markov switching are carried out. The simulation studies not only confirm the consistency, but also give valuable information on the finite sample property of the MLE, see Xie et al. [Paper V].

Theorem 3.4.1 (Xie et al. [Paper V, Theorem 1]) *For the GARMS model (24), assume (A1)-(A5). Let $\hat{\phi}_n$ be an MLE sequence over Φ^* , satisfying*

$$L_n^*(Y_n, \dots, Y_1; \hat{\phi}_n) = \sup_{\phi \in \Phi^*} L_n^*(Y_n, \dots, Y_1; \phi), \text{ a.s.}$$

then $\hat{\phi}_n$ tends to ϕ_0 a.s. as $n \rightarrow \infty$.

4 LSW processes and forecasting

Recently, many work have been done in which not only the traditional time domain techniques are used but also the time-scale or time-frequency techniques, such as wavelet transforms. One advantage of using wavelet transform is that it depends less on specifications of the dependent structure and distribution of the original series since wavelet coefficients are often less correlated than the original data. In some time-frequency models, certain non-stationary processes can also be treated in the same framework (see, e.g. Dahlhaus [20] and Mallat et al. [55]).

4.1 LSW processes

The locally stationary wavelet (LSW) process is a relatively new time-scale analysis tool proposed by Nason et al. [57]: it incorporates a class of stochastic processes based on non-decimated wavelets. It is well-known that a stationary stochastic process $X_t, t \in \mathbb{Z}$, can be written as

$$X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) d\zeta(\omega), \quad (26)$$

where $d\zeta(\omega)$ is an orthonormal increment process (Priestley [62]). The idea behind the LSW process is just to replace the set of harmonics $\{\exp(i\omega t) | \omega \in [-\pi, \pi]\}$ in (26)

with a set of non-decimated wavelets and the spectrum $A(\omega)$ by some time-varying quantities. For the definition and transforms of non-decimated wavelets, see Percival and Walden [60].

Definition 4.1.1 (Nason et al. [57]) *An LSW process is a sequence of doubly-indexed stochastic processes $\{X_{t,T}\}_{t=0,\dots,T-1}$, having the following representation in the mean-square sense*

$$X_{t,T} = \sum_{j=-J}^{-1} \sum_k \omega_{j,k;T} \psi_{j,k-t} \xi_{j,k}, \quad (27)$$

where $\xi_{j,k}$ is a random orthonormal increment sequence and where $\psi_{j,k}$ is a discrete non-decimated family of wavelets for $j = -1, -2, \dots, -J(T)$, $k = 0, \dots, T-1$ based on a mother wavelet $\psi(t)$ of compact support. The following properties are also assumed:

- (I) $E\xi_{j,k} = 0$ for all j, k . Hence $EX_{t,T} = 0$ for all t and T .
- (II) $\text{cov}(\xi_{j,k}, \xi_{l,m}) = \delta_{jl} \delta_{km}$.
- (III) The amplitudes $\omega_{j,k;T}$ are real constants and for each $j \leq -1$ there exists a Lipschitz-continuous function $W_j(z)$ for $z \in (0, 1)$ which satisfies

$$\sum_{j=-\infty}^{-1} W_j^2(z) < \infty \text{ uniformly in } z \in (0, 1)$$

with Lipschitz constants L_j which are uniformly bounded in j and

$$\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty.$$

In addition, there exists a sequence of constants C_j fulfilling $\sum_j C_j < \infty$ such that for each T

$$\sup_{k=0,\dots,T-1} |\omega_{j,k;T} - W_j(k/T)| \leq C_j/T.$$

Fryźlewicz [33] showed that by slightly altering the Assumption (III) in Definition 4.1.1, LSW processes can capture all the stylized properties of financial time series (P1-P4) described in Section 1. Note that a set of non-decimated wavelets does not constitute a basis for the underlying space, hence $\omega_{j,k}$ may not be uniquely determined. LSW processes are still meaningful by defining an Evolutionary Wavelet Spectrum (EWS) of sequence $\{X_{t,T}\}_{t=0,\dots,T-1}$ for infinite sequence $T \geq 1$ as

$$S_j(z) = W_j^2(z), \text{ for } j = -1, \dots, -J(T), z \in (0, 1).$$

Under Assumption (III) of Definition 4.1.1, $S_j(z) = \lim_{T \rightarrow \infty} |\omega_{j,[zT];T}|^2$ and $\sum_{j=-\infty}^{-1} S_j(z) < \infty$ uniformly in $z \in (0, 1)$. By realizing the covariance structure of a LSW process with lag τ as

$$\text{Cov}(X_{t,T}, X_{t+\tau,T}) = \sum_j \sum_k \omega_{j,k;T}^2 \psi_{j,k-t} \psi_{j,k-t-\tau}, \quad (28)$$

it is clear that EWS actually measures the variance at a particular time z and scale j , which is the analogue of the usual spectrum for stationary processes. Define the autocovariance as follows:

$$c_T(z, \tau) = \text{Cov}(X_{[zT], T}, X_{[zT] + \tau, T})$$

and the local autocovariance with EWS $S_j(z)$ as

$$c(z, \tau) = \sum_{j=-\infty}^{-1} S_j(z) \Psi_j(\tau),$$

where the $\Psi_j(\tau) = \sum_{k=-\infty}^{\infty} \psi_{j,k} \psi_{j,k-\tau}$ is defined as the autocorrelation wavelets. From (28) it can be seen that $\|c_T - c\|_{L_\infty} = O(T^{-1})$ (Nason et al. [57]). Recall that the local variance $\sigma^2(z) := c(z, 0) = \sum_{j=-\infty}^{-1} S_j(z)$ since $\Psi_j(0) = 1$ for all values of j . Hence, the variances and covariances of a LSW process can be estimated by estimating the EWS.

Nason et al. [57] showed that, assuming innovations ξ_t in Definition 4.1.1 are Gaussian, an unbiased estimator of the EWS vector $\mathbf{S}(k) = \{S_j(k/T)\}_{j=-1, \dots, -J}$ for the LSW process $X_{t,T}$ is

$$\mathbf{A}_J^{-1} \mathbf{I}(k),$$

where \mathbf{A}_J is the inner product of the autocorrelation wavelet, whose element $A_{j,l} = \langle \Psi_j, \Psi_l \rangle = \sum_{\tau} \Psi_j(\tau) \Psi_l(\tau)$, and $\mathbf{I}(k)$ is the vector of the wavelet periodogram, the element of which is the square of empirical wavelet coefficients defined by

$$d_{j,k;T} = \sum_{t=0}^{T-1} X_{t,T} \psi_{j,k-t}, \quad (29)$$

where $\psi_{j,k}$ is the same wavelet basis used to build $X_{t,T}$ in Definition 4.1.1. Therefore, the estimation of EWS and local variance and covariance is feasible. See Nason et al. [57] or Xie et al. [Paper IV] for further discussion about this estimation.

4.2 Wavelet basis selection

As seen in Section 4.1, in order to ensure an unbiased estimator of the EWS, the wavelet periodogram has to be constructed using the true wavelet (see (29)) as in the definition of $\{X_{t,T}\}$. However, in practice it is unrealistic to assume that we know the true wavelet. It is natural to ask how a practitioner can choose an appropriate wavelet basis on which to build the model, and what happens if an inappropriate one is chosen. Since these questions were first posed by Nason et al. [57] they have not been answered in the literature. Xie et al. [Paper IV] conducted a sensitivity analysis, based on numerical examples, to demonstrate the effect of selecting the wrong wavelet on the estimate of EWS.

The analysis was conducted as follows. A set of wavelets for the comparison, mainly the compactly supported orthogonal wavelets from Daubechies [21] with different filter lengths, was predetermined. Some true LSW processes based on these wavelets, including stationary, non-stationary ones with known EWS and local variance, were constructed. For each process, 50 realizations are generated. These wavelets were then

applied to all realizations to construct the wavelet periodogram and estimate the EWS. The estimates were then compared with the true values and the averages (over 50 realizations) of mean square errors (MSE) were obtained as a criterion for selection.

From Xie et al. [Paper IV], for the non-stationary process, a quite clear conclusion is that choosing different wavelet bases is not particularly sensitive; using the least-symmetric wavelet s8 usually gives the smallest MSE no matter which wavelet generates the process. This implies that for a non-stationary analysis based on LSW processes, a default choice could be the s8 wavelet. For a stationary process, the selection is more sensitive to the true wavelet basis. Xie et al. [Paper IV] identified a ‘cutting’ property associated with the covariance of LSW processes that may help to detect the length of the wavelet filter.

4.3 A new forecasting algorithm

Fryźlewicz et al. [35] developed a forecasting algorithm for LSW processes. The predictor for the h steps ahead forecast of $X_{t-1+h,T}$, given observations $X_{0,T}, X_{1,T}, \dots, X_{t-1,T}$, is defined as

$$\hat{X}_{t-1+h,T} = \sum_{s=0}^{t-1} b_{t-1-s;T} X_{s,T}. \quad (30)$$

The coefficients $b_{j,T}, j = 0, \dots, t-1$, are chosen to minimize the MSPE $E(\hat{X}_{t-1+h,T} - X_{t-1+h,T})^2$. That is, the vector $\mathbf{b}_t = (b_{0,T}, \dots, b_{t-1,T})^T$ is such that

$$\mathbf{b}_t = \arg \min_{\mathbf{b}_t'} \left[(\mathbf{b}_t'^T, -1) \Sigma_{t+h-1;T} (\mathbf{b}_t'^T, -1)^T \right], \quad (31)$$

where $\Sigma_{t+h-1;T}$ is the covariance matrix of $X_{0,T}, \dots, X_{t-1,T}$ and $X_{t-1+h,T}$. Directly taking the derivative over the quadratic form in (31) then equating it to zero leads to a linear equation system for solving \mathbf{b}_t

$$\Sigma_{t-1;T} \mathbf{b}_t = \mathbf{C}_{t-1+h} \triangleq \frac{\mathbf{C}_{t-1,h} + \mathbf{C}_{h,t-1}^T}{2}, \quad (32)$$

where $\Sigma_{t-1;T}$ is the covariance matrix of $X_{0,T}, \dots, X_{t-1,T}$, $\mathbf{C}_{t-1,h}$ is the column vector of covariances between $X_{0,T}, \dots, X_{t-1,T}$ and $X_{t-1+h,T}$, and $\mathbf{C}_{h,t-1}$ the vector of covariances between $X_{t-1+h,T}$ and $X_{0,T}, \dots, X_{t-1,T}$. These (co)variances can be estimated by estimating the local autocovariance.

In practice, $\Sigma_{t;T}$ in (31) can be approximated by its estimate (cf. Fryźlewicz et al. [35]). In addition, considering the non-stationary nature and local smoothness of the process, it is recommended that only the most recent p observations in (30) should be used, rather than the entire sequence, i.e.,

$$\hat{X}_{t-1+h,T}^{(p)} = \sum_{s=t-p}^{t-1} b_{t-1-s;T} X_{s,T}. \quad (33)$$

The parameter p , as well as some other parameters of the estimation, can be selected data-driven by so-called Adaptive Forecasting method (see Fryźlewicz et al. [35] for details).

Fryżlewicz's algorithm can work well for short forecasting horizons (usually small values of p 's) and carefully chosen parameters (see Fryżlewicz et al. [35] and Fryżlewicz [33] for examples). However, using this algorithm often results in an extraordinarily high value of \mathbf{b}_t when solving (32) because the covariance matrix often becomes singular, even for moderately large value t in (30) (or p in (33)). Consequently, the forecasts predict abnormally large values (outliers). Close investigation reveals that this problem is difficult to circumvent without artificially intervening on a case-by-case basis.

In order to overcome this weakness, Xie et al. [Paper IV] suggest imposing some restriction on the predictor coefficients \mathbf{b}_t when minimizing the quadratic form of (31). For instance, an obvious constraint is to require

$$\mathbf{b}_t^T \mathbf{1} = 1, \quad (34)$$

where $\mathbf{1}$ is the unit vector with the same length as \mathbf{b}_t . This actually works as a weighted average predictor with data-driven coefficients. The solution of (31) with constraint (34) is easily obtained using the Lagrangian Multiplier (LM) method via the equation system

$$\begin{pmatrix} \Sigma_{t-1;T} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{b}_t \\ -\frac{1}{2}\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{t-1+h} \\ 1 \end{pmatrix}, \quad (35)$$

where λ is the Lagrangian multiplier. However, remember that imposing constraint (34) cannot prevent the excessively large predictor coefficients from occurring. Hence, it does not fit our purpose. Another convenient choice may be more preferable, namely the requirement for a unit length of vector \mathbf{b}_t , i.e.,

$$\mathbf{b}_t^T \mathbf{b}_t = 1. \quad (36)$$

It should be mentioned that, by imposing a restriction to \mathbf{b}_t , the parameter space of \mathbf{b}_t is reduced and we may obtain only local maxima. In addition, the solving of \mathbf{b}_t under condition (36) is not so direct as in (35). Using the LM method, the solution of \mathbf{b}_t with the unit-length restriction is formally

$$\mathbf{b}_t = (\Sigma_{t-1;T} - \lambda \mathbf{I})^{-1} \mathbf{C}_{t-1+h}, \quad (37)$$

where \mathbf{I} is the identity matrix of the same size as Σ , λ is again the Lagrangian multiplier and satisfies

$$\mathbf{C}_{t-1+h}^T (\Sigma_{t-1;T} - \lambda \mathbf{I})^{-1} (\Sigma_{t-1;T} - \lambda \mathbf{I})^{-1} \mathbf{C}_{t-1+h} = 1 \quad (38)$$

and

$$\Sigma_{t-1;T} - \lambda \mathbf{I} > 0 \text{ (positive definite)}. \quad (39)$$

So (37) can be solved numerically.

The new forecasting algorithm, with constraint (36), together with Fryżlewicz's algorithm, has been applied to real data sets to conduct out-of-sample experiments. It has been shown (Xie et al. [Paper IV]) that while the performance of Fryżlewicz's algorithm is severely affected by the outliers, the new algorithm works consistently and outperforms Fryżlewicz's algorithm in most cases. See Xie et al. [Paper IV] for the comparison and a discussion of the result.

4.4 An application to volatility forecasting

The new forecasting algorithm paves the way for volatility forecasting using LSW processes. The volatility forecasts may be obtained by estimating the EWS of the sequence together with the forecasts. Fryżlewicz's algorithm is not suitable for this volatility forecasting method due to the occurrence of outliers. Wavelet s8 is used as a representative in this application and some other wavelets are also discussed.

The volatility forecasting ability of the LSW method was compared with those of the standard GARCH(1,1), GARCH(1,1) with t -distribution for innovations η_t in (8), EGARCH(1,1), and RS-GARCH(1,1) models. The forecasting was conducted on r_t , the log returns of the daily S&P500 index from 2 Jan 1990 to 29 Dec 2000, from the Center for Research in Securities Prices database. To perform the out-of-sample forecast, starting from a half sample ($t = 1390$, 29 June 1995), the parameters were estimated using all previous observations and the adaptive forecasting estimation method. All the different models were tested for 1 to 50 steps ahead forecasting. After every 50 steps, the data were updated, parameters re-estimated and forecasting conducted again. The sample MSPEs with respect to the true volatilities ($\sigma_t^{2,*}$, see equation (40) below) for all forecasting horizons were summarized as an evaluation criterion.

One difficulty associated with volatility forecasting, compared with forecasting the actual observations, is that the true volatility is unknown. There is much discussion in the literature about the definition of true volatility. Readers are referred to Andersen and Bollerslev [2], Střáicř and Granger [67], and a comprehensive review on volatility forecasting by Poon and Granger [61]. Střáicř and Granger [67] showed that a long return series of the S&P500 index is non-stationary, but exhibits a locally stationary structure. Inspired by this finding, a new true volatility definition, as a local mean of squared observations over a symmetric interval $(t - m, t + m)$ around the observation at time t for some positive integer m , is proposed by Xie et al. [Paper IV] as

$$\sigma_t^{2,*} = \frac{1}{2m+1} \sum_{i=-m}^m r_{t+i}^2 - \left(\frac{\sum_{i=-m}^m r_{t+i}}{2m+1} \right)^2. \quad (40)$$

This volatility definition can also be applied to stationary processes. However, it seems particularly suitable for processes with a locally stationary structure, smooth evolution of the variance or even variances that exhibit a linear trend. The length of the selected interval, $l = 2m + 1$, can be adjusted for different processes. A fairly large l may be used for stationary processes, and a smaller one for non-stationary processes. In Xie et al. [Paper IV], $l = 5, 11, 19$ and 31 were used, and their differences discussed.

The ratios of the MSPE of different GARCH models, over that of the LSW modeling with the s8 wavelet, are presented in Figure 3, for all forecasting horizons. The ratios for the RS-GARCH model are usually over five for most forecasting horizons and not included in the figure. Perhaps surprisingly, in-sample estimation of the data shows that only one regime is visible. Regime prediction also always adheres to a single regime. In this case, the model is too complicated to produce an accurate forecast. Figure 3 suggests that the volatility forecasting using LSW processes performs fairly well; LSW forecasting usually produces more accurate forecasts for most forecasting horizons. In particular, it outperforms GARCH models for smaller horizons, where GARCH models are known to be capable of giving good forecasting. For more discussions of the

comparison between different interval lengths and wavelet bases, see Xie et al. [Paper IV].

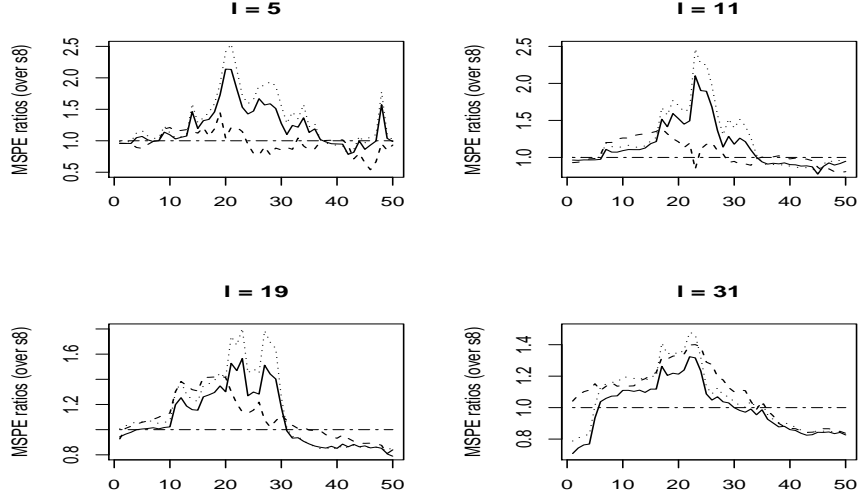


Figure 3: The ratios of GARCH(1,1) (solid), EGARCH(1,1) (dashed) and GARCH-t(1,1) (dotted) MSPEs divided by corresponding MSPE from the LSW modeling with s8 wavelet basis in forecasting S&P500 return series, and unit line (dash-dotted), against the forecasting horizons. The lengths of intervals (l) in (40) are 5, 11, 19, and 31, respectively.

5 Summary of the papers

In this section, a brief summary of the papers is presented. Please refer to the appended papers for details.

Paper I. In this paper, we investigate the asymptotic properties of the quasi-maximum likelihood estimator for the GARCH(1,2) model under stationary innovations. Consistency of the global QMLE and asymptotic normality are obtained, which extends the previous results for GARCH(1,1) by Lee and Hansen [47] under weaker conditions.

Paper II. The regime-switching GARCH model combines the idea of Markov switching and the GARCH model, and also includes the popular Hidden Markov models as special cases. The statistical inference associated with this model, however, is rather difficult because the observations depend on the whole regime path due to the recursive structure of the GARCH equation. In this paper, inspired by the work of Gray [37] on the GARCH(1,1) model, we consider a reduced regime-switching GARCH(p, q) model, that is, the past regimes are integrated out at every step and observations then depend only on the current regimes. The Maximum Likelihood (ML) estimation is considered and the consistency of ML estimators for this model is proved. Sim-

ulation studies to illustrate the consistency and asymptotic normality of the proposed estimators are presented. In a model specification problem, where an ordinary GARCH model is wrongly specified to data generated from regime-switching GARCH models, the persistence of model is discussed; the finding is interesting.

- Paper III. The regime-switching GARCH model incorporates the idea of regime switching into the more restrictive GARCH model, which significantly extends the GARCH model. However, the statistical inference for such an extended model is rather difficult because observations at any time point then depend on the whole regime path and the likelihood quickly becomes intractable as the length of observations increases. In this paper, by transforming it to an infinite order ARCH model, we are able to derive a likelihood that can be handled directly and the consistency of the maximum likelihood estimators is proved. Simulation studies illustrate the consistency and asymptotic normality of the estimators. Both Gaussian and non-Gaussian innovations are investigated. A model specification problem is also presented.
- Paper IV. Locally stationary wavelet (LSW) processes, built on non-decimated wavelets, can be used to analyze and forecast non-stationary time series, and they have been proved useful in the analysis of financial data. In this paper we first carry out a sensitivity analysis using numerical examples, and propose some practical guidelines for choosing the wavelet bases for these processes. The existing forecasting algorithm from Fryżlewicz et al. [35] is found to have no protection from outliers and a new algorithm, imposing restrictions on the predictor coefficients, is proposed. These algorithms are tested on real data and the new algorithm works consistently and outperforms the existing algorithm in most of the cases. The volatility forecasting ability of LSW modeling based on our new algorithm is then discussed and is shown to be competitive against traditional GARCH models when applied to S&P500 return series. The applications in Nason et al. [57] and Fryżlewicz et al. [35] are limited to the Haar wavelet. In this paper, this limitation is relaxed and many others wavelets are applied.
- Paper V. In this paper, a general autoregressive model with Markov switching is considered, where the autoregression may be of an infinite order. The consistency of the maximum likelihood estimators for this model is obtained. The assumptions for consistency are discussed in some detail. The conditions associated with consistency are discussed using finite and infinite order Markov switching AR models. The simulation studies using these examples demonstrate the assumptions and the consistency of the MLEs. The asymptotic normality is also suggested by numerical experiments.

6 Future research

There are many topics for potential future research suggested by the results of this thesis. Here I outline and discuss some of them from the perspectives of theoretical development and empirical applications.

6.1 Theoretical development

As discussed in Sections 3.2 and 3.3, the asymptotic normality of MLEs for RS-GARCH models is still an open problem. Bickel et al. [6] provided a good reference for the asymptotic normality for HMMs. Their approach is the classic Cramér method, involving a central limit theorem for the score function and a law of large numbers for the observed information. However, they require an inequality ([6, Lemma 3]), which can be traced back to a 1966 paper by Baum and Petrie [4] on the HMM with finite states and observations taking values in a finite set. The property of conditional independence of observations given their corresponding regimes is crucial for this inequality. This property does not hold for Markov switching models and a suitable condition to circumvent the inequality is not apparent. Relatively, Douc et al.'s result [24] may be closer to what we want. Douc et al. obtained the consistency and asymptotic normality for MLE of the ARMS (see Section 3.4) with a continuous state-space Markov chain, which includes the RS-ARCH model. However, their results require the Markov property implied by the ARCH model to derive an important inequality ([24, Lemma 1]), which is not valid for the GARCH structure. The asymptotic normality for RS-GARCH models is challenging, but also attractive. It requires more subtle techniques. Perhaps some 'truncation' technique applying to the RS-ARCH(∞) model and some particular case (e.g. a finite state Markov chain) could be a good starting point.

In Paper I, only the asymptotics for the GARCH(1,2) model (under dependent innovations) are considered. In fact, a direct extension to GARCH(1, q) requires barely more complexity of denotations. For a general GARCH(p , q) model, investigation reveals that the uniqueness of the ARCH representation of the GARCH model of Berkes et al. [5] relies on the i.i.d. assumption of innovations. Otherwise, Berkes et al.'s proofs mimic what Lee and Hansen [47] showed. Therefore, it is possible to extend our result in Paper I further to GARCH(p , q) model if we can work around the unique representation. For the GARMS model (24), an attractive setting would be an infinite-dimensional parameter space since the autoregression can be of infinite order. However, it is difficult even to define a distance for an infinite-dimensional space that achieves the compactness of the parameter space and consistency of the estimators simultaneously. Another potential generalization of (24) is to relax the i.i.d. innovation $\{\varepsilon_t\}$ to certain kind of stationary process. Such a generalization would obtain the possibility to study the ML estimations of ARMA models with Markov switching, see e.g. Billio et al. [7]. These remain topics for future research.

For the reduced RS-GARCH model, the process Y_t is assumed to be strictly stationary and ergodic (see Paper II and Section 3.2) for the consistency of the MLE. However, the condition that ensures the stationarity of Y_t is not clear. By writing the RS-GARCH model (18) as a general autoregressive model with stationary and ergodic coefficients, i.e. equation (22), Francq et al. [32] obtained such a condition for the RS-GARCH model, i.e. (23), the top Lyapunov exponent ρ^* is strictly negative. This result is stimulating, however a direct analogue does not exist for the reduced RS-GARCH model. More work needs to be undertaken. It should be noted that, for many of the extensions of the GARCH model mentioned in Section 2.5, the basic properties such as conditions for stationarity, existence of moments, consistency, and asymptotic normality are also not well developed. See the review by Li et al. [49] for some available results.

Time-frequency or time-scale analysis has become more and more popular for both financial time series and in other fields (cf. Vidakovic [73]). Besides the LSW modeling, there are many other possible methods and applications using wavelets. For example, Fryźlewicz et al. [34] used (*inter alia*) Haar wavelet shrinkage to study processes with stepwise variance. We suggest that such a technique may be extended to other wavelets and may also be applied to other stochastic processes.

6.2 Empirical applications

The results for the RS-GARCH model make it possible to envisage its use in many empirical applications. We can apply the model to different financial time series to see if it describes the data better than the ordinary GARCH. It is believed that the strong persistence of the ordinary GARCH processes that are frequently observed (e.g. Cai [17] and Hamilton and Susmel [41]) can be explained by changes of regimes, see, e.g. Gray [37] and Hamilton and Susmel [41]. Among others, Hamilton and Susmel [41] showed that RS-ARCH models provided better model interpretation than ordinary GARCH; Using an adjusted RS-GARCH(1,1) model from Gray [37], Klaassen [44] showed improved volatility forecasts against the ordinary GARCH(1,1) model. The potential for applying our results is significant.

In addition, the time point of regime-switching reveals much information which is of interest to economists and practitioners. For example, Hamilton [38] used a Markov switching AR model to analyze the US GNP growth rate series and labeled the two regimes in his model as the ‘expansion’ and ‘recession’ periods. Hence, it is interesting to examine when a recession really begins. The starting months of different recessions in US identified by Hamilton [38] are very close to the official data. Such an example represents many possible interesting applications.

Methods using LSW processes or other time-frequency techniques have been proven appealing for the analysis of time series with particular spectral features, see e.g. Nason et al. [57] for an analysis of an infant heart rate (ECG) sequence. Fryźlewicz [33] (among others) showed that LSW framework is useful in financial log return series analysis. Struzik [68] applied the wavelet based effective Hölder exponent to uncover the local scaling (correlation) characteristics of the S&P index and discover structure through the analysis of collective properties of non-stationary behaviour. Therefore, not only the stationary, but also non-stationary processes can be analyzed by the wavelet transform or other time-frequency methods (see e.g. Dahlhaus [20], Fryźlewicz et al. [35], and Xie et al. [Paper IV]). Applications in [33], [35], and [57] using LSW processes are limited to the Haar wavelet, while our work [Paper IV] extends the algorithm to many other wavelets. More applications using different wavelets to both stationary and non-stationary processes can be expected.

We should mention that, according to our simulation studies, a consistent estimate for the RS-GARCH models requires the size of observations to be as large as 3000, or even more; such data may not always be available. One possible solution in such cases is to use some resampling method to reduce estimate biases, such as the bootstrap methods, see e.g. Davison and Hinkley [22]. Such possibility has to be explored in the future too.

References

- [1] Aldrich, J. 1997. R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science* 3, 162-176.
- [2] Andersen, T.G. and Bollerslev, T. 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39, 885-905.
- [3] Baillie, R.T. 1996. Long memory processes and fractional integration in econometrics. *Journal of Econometrics* 73, 5-59.
- [4] Baum, L.E. and Petrie, T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics* 37, 1554-1563.
- [5] Berkes, I., Horváth, L. and Kokoszka, P. 2003. GARCH processes: structure and estimation. *Bernoulli* 9, 201-227.
- [6] Bickel, P.J., Ritov, Y. and Rydén, T. 1998. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics* 26, 1614-1635.
- [7] Billio, M., Monfort, A. and Robert, C.P. 1999. Bayesian estimation of switching ARMA models. *Journal of Econometrics* 93, 229-255.
- [8] Bollerslev, T. 1986. Generalized autoregressive heteroscedasticity. *Journal of Econometrics* 31, 307-327.
- [9] Bollerslev, T. 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach. *Review of Economics and Statistics* 72, 498-505.
- [10] Bollerslev, T., Chou, R.Y. and Kroner, K.F. 1992. ARCH modeling in finance. A review of the theory and empirical evidence. *Journal of Econometrics* 52, 5-59.
- [11] Bollerslev, T., Engle, R.F. and Wooldridge, J.M. 1988. A capital asset pricing model with time varying covariances. *Journal of Political Economy* 96, 116-131.
- [12] Bougerol, P. and Picard, N. 1992. Strict stationarity of generalized autoregressive processes. *The Annals of Probability* 20, 1714-1729.
- [13] Bougerol, P. and Picard, N. 1992. Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* 52, 115-127.
- [14] Box, G.E.P. and Jenkins, G.M. 1990. *Time Series Analysis: Forecasting and Control*. Revised edition. Holden-Day: San Francisco.
- [15] Brockwell, P., Chandraa, E. and Lindner, A. 2006. Continuous-time GARCH processes. *Annals of Applied Probability* 16, 790-826.
- [16] Brockwell, P. and Davis, R.A. 1991. *Time Series: Theory and Methods*. Second Edition. Springer-Verlag: New York.

- [17] Cai, J. 1994. A Markov model of switching-regime ARCH. *Journal of Business and Economic Statistics* 12, 309-316.
- [18] Cont, R. 2001. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* 1, 223-236.
- [19] Cox, D.R. 2006. *Principles of Statistical Inference*. Cambridge University Press: Cambridge.
- [20] Dahlhaus, R. 1997. Fitting time series models to nonstationary processes. *The Annals of Statistics* 25, 1-37.
- [21] Daubechies, I. 1992. *Ten Lectures on Wavelets*. SIAM: Philadelphia.
- [22] Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap Methods and Their Application*. Cambridge University Press: Cambridge.
- [23] Ding, Z., Granger C.W.J. and Engle R.F. 1993. A long-memory property of stock market returns and a new model. *Journal of Empirical Finance* 1, 107-131.
- [24] Douc, R., Moulines, É. and Rydén, T. 2004. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics* 32, 2254-2304.
- [25] Engle, R.F. 1982. Autoregressive conditional heteroscedasticity with estimates of U.K. inflation. *Econometrica* 50, 987-1008.
- [26] Engle, R.F. and Bollerslev, T. 1986. Modelling the persistence of conditional variances. *Econometric Reviews* 5, 1-50, 81-87.
- [27] Engle, R.F., Lilien, D.M., and Robins, R.P. 1987. Estimating time varying risk premia in the term structure: The Arch-M model. *Econometrica* 55, 391-407.
- [28] Fama, E.F. 1965. The behavior of stock market prices. *Journal of Business* 38, 34-105.
- [29] Fan, J.Q. 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag: New York.
- [30] Francq, C. and Roussignol, M. 1998. Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator. *Statistics* 32, 151-173.
- [31] Francq, C. and Zakoïan, J. 2004. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* 10, 605-637.
- [32] Francq, C., Roussignol, M. and Zakoïan, J. 2001. Conditional heteroscedasticity driven by hidden Markov chains. *Journal of Time Series Analysis* 22, 197-220.
- [33] Fryźlewicz, P. 2005. Modeling and forecasting financial log-returns as locally stationary wavelet processes. *Journal of Applied Statistics* 32, 503-528.
- [34] Fryźlewicz, P., Sapatinas, T. and Subba Rao, S. 2006. A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika* 93, 687-704.

- [35] Fryzlewicz, P., Van Bellegem, S. and von Sachs, R. 2003. Forecasting non-stationary time series by wavelet process modeling. *The Annals of Institute of Statistical Mathematics* 55, 737-764.
- [36] Glosten, L., Jagannathan R. and Runkle, D. 1993. Relationship between the expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance* 48, 1779-1801.
- [37] Gray, S.F. 1996. Modelling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* 42, 27-62.
- [38] Hamilton, J.D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357-384.
- [39] Hamilton, J.D. 1994. *Time Series Analysis*. Princeton University Press: New Jersey.
- [40] Hamilton, J.D. and Raj, B. 2002. New directions in business cycle research and financial analysis. *Empirical Economics* 27, 149-162.
- [41] Hamilton, J.D. and Susmel, R. 1994. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics* 64, 307-333.
- [42] He, C. and Teräsvirta, T. 1999. Fourth moment structure of the GARCH(p , q) process. *Econometric Theory* 15, 824-846.
- [43] Heyde, C.C. 1997. *Quasi-Likelihood and its Application. A General Approach to Optimal Parameter Estimation*. Springer: New York.
- [44] Klaassen, F. 2002. Improving GARCH volatility forecasts with regime-switching GARCH. *Journal of Empirical Economics* 27, 363-394.
- [45] Klüppelberg, C., Lindner, A. and Maller, R. 2004. A continuous time GARCH process driven by a Lévy process: stationarity and second order behavior. *Journal of Applied Probability* 44, 601-622.
- [46] Krishnamurthy, V. and Rydén, T. 1998. Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of Time Series Analysis* 19, 291-307.
- [47] Lee, S. W. and Hansen, B. E. 1994. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory* 10, 29-52.
- [48] Leroux, B.G. 1992. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and Their Applications* 40, 127-143.
- [49] Li, W.K., Ling, S. and McAleer, M. 2002. Recent theoretical results for time series models with GARCH errors. *Journal of Economic Survey* 16, 245-269.
- [50] Ling, S. and Li, W.K. 1997. On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. *Journal of American Statistic Association* 92, 1184-1194.

- [51] Ling, S. and McAleer, M. 2002. Necessary and sufficient moment conditions for the GARCH(r, s) and asymmetric power GARCH(r, s) models. *Econometric Theory* 18, 722-729.
- [52] Ling, S. and McAleer, M. 2003. Asymptotic theory for a vector ARMA-GARCH model. *Econometric Theory* 19, 280-310.
- [53] Lumsdaine, R. L. 1991. *Asymptotic Properties of the Quasi-Maximum Likelihood Estimator in GARCH(1,1) and IGARCH(1,1) Models*. Dissertation, Princeton University.
- [54] MacDonald, I.L. and Zucchini, W. 1997. *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman and Hall: London.
- [55] Mallat, S.G., Papanicolaou, G. and Zhang, Z. 1998. Adaptive covariance estimation of locally stationary processes. *The Annals of Statistics* 26, 1-47.
- [56] McNees, H. 1979. The forecasting record for the 1970's. *New England Economic Review*, Sep/Oct, 33-53.
- [57] Nason, G.P., von Sachs, R. and Kroisandt, G. 2000. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of Royal Statistic Society B*, 62, 271-292.
- [58] Nelson, D. 1990. Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory* 6, 318-334.
- [59] Nelson, D. 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59, 347-370.
- [60] Percival, D.B. and Walden, A.T. 2000. *Wavelet Methods for Time Series Analysis*. Cambridge University Press: Cambridge.
- [61] Poon, S.-H. and Granger, C.W.J. 2003. Forecasting volatility in financial markets: A review. *Journal of Economic Literature* XLI, 478-539.
- [62] Priestley, M.B. 1981. *Spectral Analysis and Time Series*. Academic Press: London.
- [63] Sentana, E. 1995. Quadratic ARCH models. *Review of Economic Studies* 62, 639-661.
- [64] Shephard, N. 1996. Statistical aspects of ARCH and stochastic volatility. In D.R. Cox, D. Hinkley, and O.E. Barndorff-Nielsen (Eds.) *Likelihood, Time Series with Econometric and Other Applications*. Chapman and Hall: London.
- [65] Straumann, D. 2005. *Estimation in Conditionally Heteroscedastic Time Series Models*, In P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger (eds.) *Lecture Notes in Statistics 181*. Springer: Berlin, Heidelberg, New York.
- [66] Straumann, D. and Mikosch, T. 2006. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *The Annals of Statistics* 34, 2449-2495.

- [67] Stărică, C. and Granger, C.W.J. 2005. Nonstationarities in stock returns. *Reviews of Economic Statistics* 87, 503-522.
- [68] Struzik, Z.R. 2001. Wavelet method in financial time-series processing. *Physica A* 296, 307-319.
- [69] Taylor, S. 1986. *Modelling Financial Time Series*. Wiley: New York.
- [70] Teräsvirta, T. 2006. An introduction to univariate GARCH models. Working Paper Series in Economics and Finance 646, Stockholm School of Economics. Forthcoming in T.G. Andersen, R.A. Davis, J.-p. Kreiss and T. Mikosch (eds.) *Handbook of Financial Time Series*, Springer-Verlag: New York.
- [71] Wedderburn, R.W.M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61, 439-447.
- [72] Weiss, A. A. 1986. Asymptotic theory for ARCH models: Estimation and testing. *Econometric Theory* 2, 107-131.
- [73] Vidakovic, B. 1999. *Statistical Modeling by Wavelets*. Wiley: New York.
- [74] Yao, J. and Attali, J.-G. 2000. On stability of nonlinear AR processes with Markov switching. *Advances in Applied Probability* 32, 394-407.
- [75] Zakoian, J.M. 1994. Threshold heteroskedastic functions. *Journal of Economic Dynamics and Control* 18, 931-955.

Acknowledgements

First of all, I want to express my deep gratitude to my two supervisors Associate Professor Jun Yu and Professor Bo Ranneby, for providing me the opportunity to study in SLU, a flexible working environment, and the support all the way. The thesis would not exist without your scientific knowledge and unselfish devotion, in particular the enormous efforts and time-spending during the preparation of this thesis. Also thank you, Bosse, for always being available for help! Thank you, Jun, for all the help, from scientific thinking to Swedish and English, from researches to badminton-playing, from computer stuffs to my family. Thank you, you are Super Advisors!

Many other should be included in my thank-you list. Associate Professors Magnus Ekström and Peichen Gong carefully read a draft of the thesis and articles, and provided me many valuable comments. I should also thank Magnus for teaching me the HPC2N things that have facilitated the computations, and even HPC2N in Umeå University should be covered by the thanks. I learned a lot from Professor Yuri Belyaev, now and then, not only of this thesis, but also of the articles and general scientific ideas. Professor Bengt Kriström always gives me encouragement, different ideas from economic side, and help. I appreciate very much the interesting conversations during the late lunches with Professor Peter Lohmander, including the discussions about the thesis and my articles, and many other topics. Peichen, Ola, Solveig, Wenchao, Fadian, Camilla, Cecilia and other colleagues in our department also helped me in one way or another, with the thesis or a daily life. Thank you, all! My Chinese friends in Umeå deserve a position in this list too; you make my life in Umeå more enjoyable.

To my parents, I owe them a lot. I just hope I can make up later. My in-laws, especially my mother-in-law, have been helpful for us these years. Now, finally, my thanks go to my dear wife, Xiaohua. Thank you for all your love, support, and accompanying me along all these time, be it good or bad, happy or sad. Also thank you for bringing us our lovely and beloved, of course sometimes naughty daughter, Xixi! I believe then, I don't need to worry that I will have nothing to do in my spare time.